Statistique et société
juin 2015

Volume 3, Numéro 1

Deux débats sur les données

www.statistique-et-societe.fr

Sommaire
Statistique et société

Volume 3, Numéro 1

Sommaire
Statistique et Société

Volume 3, Numéro 1

Varia

Statistique et société

Magazine trimestriel publié par la Société Française de Statistique.
Le but de Statistique et société est de montrer d'une manière attrayante et qui invite à la réflexion l'utilisation pratique de la statistique dans tous les domaines de la vie, et de montrer comment l'usage de la statistique intervient dans la société pour y jouer un rôle souvent inaperçu de transformation, et est en retour influencé par elle. Un autre dessein de Statistique et société est d'informer ses lecteurs avec un souci pédagogique à propos d'applications innovantes, de développements théoriques importants, de problèmes actuels affectant les statisticiens, et d'évolutions dans les rôles joués par les statisticiens et l'usage de statistiques dans la vie de la société.

Design graphique
fastboil.net

ISSN 2269-0271

Éditorial

Emmanuel DIDIER

Rédacteur en chef de Statistique et société

Chers lecteurs,

Décidément, notre revue donne à chaque livraison un peu plus raison au mot d'ordre d'Alain Desrosières lorsqu'il invitait à « discuter l'indiscutable ». Il signifiait ainsi que, d'un côté, le profane appréhende la statistique comme étant indiscutable parce que quantitative et scientifique. Mais d'un autre côté, ce même profane observe qu'elle participe chaque jour un peu plus au débat public et souhaite donc à juste titre s'en emparer pour la discuter, la contrebalancer, l'évaluer.

En lisant le titre des textes rassemblés dans le présent numéro de *Statistique et société*, on tombe sur les mots de « débat », « défi », « manifeste », « communication » ... et enfin, comme d'habitude, sur un café de la statistique. Débat sur l'accès aux données de santé ; défi des Big Data ; manifeste proposant de remanier la nomenclature des PCS ; communication plaidant pour une recherche plus transparente ; et enfin un café statistique où les conversations ont porté sur les nouveaux cadres de la validation des recherches scientifiques. Au delà des dossiers auxquels nous nous étions habitués, ce sont autant de formats d'échange dans lesquels la quantification, sous ses aspects les plus contemporains, se trouve effectivement rendue discutable.

Le débat sur l'accès aux données fait suite à un précédent numéro consacré à cette question. Merci aux auteurs d'avoir pris ou repris la plume pour partager avec nous leur réaction. N'hésitez pas, chers lecteurs, à reproduire ce geste. Preuve est faite que nous prenons vos retours très au sérieux !

Bonne lecture

Débat sur l'accès aux données de santé

Suite au dossier paru dans Statistique et société vol.2 n°2 - mai 2014

Avec la participation de
Nicolas Belorgey, sociologue,
Jean-Pierre Le Gléau, inspecteur général honoraire de l'Insee,
Marcel Goldberg et Marie Zins, épidémiologistes.

Les modalités d'accès aux bases de données individuelles administratives, dès lors qu'elles sont devenues techniquement possibles, constituent un sujet de débat sociétal sensible, abordé, à travers le prisme particulier des données de santé, dans le numéro 2014- 2 de notre revue[1]. Nicolas Belorgey, chargé de recherche en sociologie au CNRS, a adressé à la rédaction de Statistique et Société, en réaction, une contribution intitulée « *Pour un véritable accès aux données de santé* » soulignant des particularités de la recherche en sociologie, que les dispositifs décrits, existants ou proposés, n'apparaissent pas avoir pris en compte, et pour lesquelles l'auteur formule quelques propositions. Nous avons demandé aux auteurs les plus concernés du numéro 2014-2 de réagir sur le texte de Nicolas Belorgey : Jean-Pierre Le Gléau, inspecteur général honoraire de l'Insee, qui suggérait dans son article de l'an dernier que le dispositif de centre d'accès sécurisé (CASD) mis en place et fonctionnel pour des recherches, études et évaluations de données économiques sensibles, pourrait inspirer un dispositif, à ce jour limité, concernant l'accès aux données de santé ; et Marcel Goldberg et Marie Zins, épidémiologistes, qui avaient exposé l'état actuel et les avancées possibles d'accès aux bases de données médico-sociales.

Sous le titre « Créer les conditions d'un accès ouvert aux données de santé », l'article 47 de la loi de santé discutée au Parlement ce printemps 2015 marque un véritable tournant en ajoutant un titre VI au code de Santé Publique, intitulé « *mise à disposition des données de santé* » dans lequel apparaît le cadre d'un système national des données de santé (SNDS) avec ses modalités d'accès. Nous espérons que le débat publié ici contribuera de façon positive à la discussion parlementaire. Statistique et société reviendra sur les principales innovations apportées par le texte de loi dans sa version définitive, après examen par les sénateurs.

1. Consultable à l'adresse : http://publications-sfds.fr/index.php/stat_soc/issue/view/36

Pour un véritable accès aux données de santé

Nicolas BELORGEY
Sociologue, CNRS

Dans son numéro 2 de 2014, la revue *Statistique et société* soulève la question de l'accès aux données de santé. Elle souligne notamment qu'en-dehors de cercles restreints à la définition peu claire, cet accès demeure trop limité, notamment dans une perspective scientifique et citoyenne (Bar Hen). Elle rappelle également que « l'Initiative transparence santé » (ITS) provient essentiellement d'acteurs du secteur de l'assurance et des produits de santé, qui ont un intérêt financier à cette ouverture (Briatte et Goeta). Or, cet intérêt n'est pas nécessairement compatible avec la préservation des droits des personnes. En effet, deux types de droits doivent être conciliés en matière d'accès aux données de santé, conciliation qui a fait l'objet de différentes lois, dont celle de 1978, dite « Informatique et libertés » : celui des personnes, qui inclut la protection de leur vie privée et, en matière de santé, la confidentialité de leur état de santé pour éviter tout profilage des assurés dans un but lucratif ; celui de l'accès à ces informations, dans un objectif d'accroissement des connaissances comme bien public.

Nous défendons ici l'idée qu'en dépit de propositions tendant à améliorer l'accès aux données de santé tout en préservant les droits des personnes, le numéro 2 de Statistique et société ne prend pas totalement en compte les besoins d'une démarche rigoureuse aujourd'hui en sciences sociales, et esquissons quelques propositions en la matière.

I. Les réflexions de ce numéro ne prennent pas totalement en compte les besoins d'une démarche rigoureuse en sciences sociales.

Ce numéro présente un certain nombre de dispositifs existants dans la statistique publique, qui permettent de concilier protection de l'anonymat des personnes et accès aux informations (Le Gléau). Ils sont utilisés notamment en matière de statistique des entreprises ou de fiscalité. Le plus perfectionné d'entre eux semble résider dans la combinaison de l'accord préalable du Comité du secret statistique (CSS) et du recours à une technique d'accès particulière, le Centre d'accès sécurisé aux données (CASD). Cette dernière permet aux chercheurs de produire des statistiques agrégées à partir des données détaillées des bases, sans pour autant pouvoir individualiser celles-ci dans le compte-rendu qu'ils en font. La proposition faite en matière d'accès aux données de santé sur la réplication de ce dispositif consiste pour les enquêteurs en « un engagement de (…) ne tenter en aucun cas d'identifier un individu précis de la base » (p. 32).

Pour progressiste qu'elle soit par rapport à la situation actuelle, cette proposition semble cependant informée par une conception fortement quantitativiste de la démarche de connaissance en sciences sociales. Or celle-ci doit être appuyée sur des outils non seulement quantitatifs, mais aussi qualitatifs. Bien sûr, l'agrégation des résultats au niveau collectif présente l'avantage de la généralité, de la plus grande extension possible des propositions avancées. Sont ainsi évités les biais de sélection que peuvent présenter les cas individuels. Mais une bonne intelligence des processus à l'œuvre suppose aussi de raisonner sur des cas individuels, comme le font les historiens et les ethnographes (M. Weber 1917; Passeron 1991; F. Weber 2009). On gagne en compréhension ce qu'on perd en extension, et *vice versa* si on passe du qualitatif pur au quantitatif pur. De fait, dans une démarche totalement rigoureuse, on allie les deux outils : raisonnement sur des cas afin d'obtenir une bonne compréhension des phénomènes au niveau le plus fin possible, tests d'hypothèses statistiques afin de vérifier le degré de généralité de ces éléments intermédiaires (Gramain et Weber 2001). En pratique, on effectue souvent des allers-retours entre ces deux niveaux, sans que l'un ait nécessairement la prééminence chronologique, méthodologique ni surtout ontologique (en matière d'interprétation du réel). Cette double démarche semble d'autant plus nécessaire que les données statistiques ont elles aussi leurs biais de construction, puisque fréquemment produites par des administrations dont elles ne sont qu'un sous produit de l'activité, comme la police (Bruno et Didier 2013) ou les hôpitaux (Belorgey 2010). Pour reprendre les termes d'Alain Desrosières (2005), on n'aurait à la limite le choix, si on ne se fiait qu'à l'une des deux méthodes prises isolément, qu'entre « décrire l'État ou explorer la société. »

En matière de données de santé, cette alliance entre approches d'origines quantitative et qualitative semble particulièrement nécessaire. En effet, les principales bases de données existantes ne sont pas conçues à des fins de recherche, donc d'objectivité scientifique, mais de gestion de services de soin (informatique hospitalière), de comptabilité (idem), ou d'administration de la Sécurité sociale (systèmes d'information de l'assurance-maladie, dont le Sniiram est la partie émergée de l'iceberg). Il y a donc fort à parier que l'écart entre l'image des personnes qui ressort de ces données et ce qu'elles sont et font en réalité, est aussi important que celui qui existe entre un relevé de dépenses de consommation de « loisir » et ce que les personnes retirent de ce type d'activité, ainsi que les déterminants de celles-ci. Les données statistiques de santé sont aujourd'hui produites à des fins thérapeutiques, organisationnelles ou d'administration du système de soin ; elles présentent donc un biais de construction important que seules des enquêtes qualitatives individualisées auprès des personnes concernées, et portant le point de vue, partiel aussi mais nécessaire à la compréhension d'ensemble, de celles-ci, peuvent apporter. Par exemple, la plupart des statistiques de santé, pour précises qu'elles puissent être sur les consommations de soin des personnes, demeurent muettes sur les appartenances sociales de celles-ci : leur métier précis, leur origine sociale, leur environnement familial élargi, etc. Ces données ne sont parfois qu'approchées par certaines variables intermédiaires, comme l'affiliation éventuelle à l'Aide médicale d'État. En leur absence, comment observer les effets de décisions publiques sur les différents groupes de la population ? Comment observer les évolutions des inégalités sociales de santé ? Etc. Pour répondre à de telles questions, il est nécessaire de recouper les données entre les bases et de recourir à des enquêtes directes auprès des personnes. Des démarches comme le projet Monaco[1] ou la cohorte Constances[2] vont dans le bon sens mais demeurent partielles.

En matière de protection des droits des personnes, la démarche par cas a depuis longtemps élaboré une réflexion et des règles en la matière. Il s'agit, là aussi, de concilier d'une part respect de la vie privée des personnes et d'autre part précision scientifique et caractère de bien public de la connaissance (Beaud et Weber 1997, chap. 9). Cette configuration s'est enrichie

1. http://www.irdes.fr/recherche/partenariats/monaco-methodes-outils-et-normes-pour-la-mise-en-commun-de-donnees-de-l-assurance-complementaire-et-obligatoire/actualites.html
2. http://www.constances.fr

récemment de la nécessité de protéger les enquêteurs des pressions dont ils peuvent être l'objet, notamment économiques et politiques (Laurens et Neyrat 2010), et les enquêtés d'une forme de spoliation de leur contribution à une œuvre commune (Weber 2011). Mais elle demeure valable : en échange de l'accès à des informations personnelles, les enquêteurs protègent les enquêtés en leur garantissant anonymat et/ou confidentialité (Beliard et Eideliman 2008). L'anonymat consiste en la simple modification des caractéristiques directement identifiantes comme le nom. La confidentialité va plus loin, puisqu'elle fait que les personnes ne puissent être identifiées même dans le milieu d'interconnaissance où elles évoluent, et suppose que des caractéristiques plus spécifiques soient brouillées, comme la présence d'un enfant handicapé dans une configuration familiale originale. Ceci peut être fait sans porter atteinte à la rigueur scientifique du résultat par différentes méthodes, comme la construction de cas fictifs pour la présentation des résultats, présentant toutes les caractéristiques essentielles du cas réel sous-jacent, mais modifié sur toutes ses caractéristiques secondaires.

Une démarche scientifique rigoureuse et protectrice des droits des personnes en matière de santé se doit donc de recouper un maximum d'informations, tant à partir de bases de données statistiques, qu'au niveau individuel, tout en garantissant anonymat voire confidentialité lors de la publication des résultats. Par rapport à la situation actuelle, cela revient à élargir, sous contrôle, la « bulle » de confinement des informations aux chercheurs (Le Gléau, page 29), comme cela a été fait en 1984 au profit des informations sur les entreprises. La logique de cette opération est en fait que ces chercheurs sont précisément les seuls à même de pouvoir recouper toutes ces informations pour faire progresser le niveau de connaissance collective.

II. Propositions pour l'accès aux données de santé

Il est donc nécessaire d'accéder aux données statistiques individuelles, ou granulaires, en amont de toute forme d'agrégation ou d'anonymisation. On peut distinguer deux cas de figure, selon que l'enquête commence par des cas concrets ou par l'accès aux données statistiques.

Dans le premier cas, un accord préalable écrit ne peut être demandé aux intéressés. Cette formule est en effet inadaptée à l'enquête en sciences sociales d'origine qualitative. Bien que codifiée dans les chartes de nombreuses associations anglo-américaines de recherche en sciences de la nature voire en sciences sociales, elle demeure en leur sein contestée, et est en fait née des abus des sciences médicales expérimentales (Bosa 2008). Par exemple, des recherches cliniques avaient été faites sans que le groupe témoin bénéficie des avantages thérapeutiques du groupe test, ce qui s'était traduit par le décès d'une grande partie de ses membres. En sciences sociales, cette visée directement thérapeutique étant par définition absente, le schéma n'est pas le même. De plus, le recueil préalable et formel d'un consentement empêche l'établissement d'une relation d'enquête. Les chercheurs américains soumis à ce type de restrictions sont parfois conduits à se positionner pour accéder aux données dans la catégorie juridique des journalistes, qui sont paradoxalement beaucoup plus libres dans leur relation à leurs enquêtés. L'enquête d'origine qualitative procède donc par des contacts personnels, qui pourraient ensuite être retrouvés dans les bases de données. Dans cette situation, l'accord peut-être beaucoup plus informel.

Le second cas de figure fonctionne en sens inverse : quand il est souhaitable de compléter les données statistiques sur certaines personnes par une enquête qualitative auprès d'elles. Comme c'est le cas aujourd'hui pour ce type de démarche, ces personnes pourraient être contactées par courrier pour recueillir leur accord préalable et écrit à toute investigation plus poussée. Un refus ou une absence de réponse de leur part vaudrait bien sûr abandon de tout approfondissement et ne serait consigné nulle part. Si elles sont d'accord, un volet qualitatif pourrait se déployer.

Dans les deux cas de figure, que l'on commence par l'étape qualitative ou quantitative, tous les recoupements d'informations seraient autorisés, et les droits des personnes enquêtées seraient protégés par le passage à l'anonymat voire à la confidentialité avant toute publication ou communication à des tiers (comme les organismes producteurs des données statistiques). Ainsi, en résumé, l'accès aux données individuelles brutes se ferait sous les conditions suivantes :

- que les cas individuels soient uniquement utilisés à des fins de compréhension des mécanismes à l'œuvre ;
- que les cas individuels finalement publiés soient anonymisés et rendus confidentiels, ce qui garantirait le respect des droits des personnes ;
- que cette présentation ne distorde pas les résultats issus du recoupement des données statistiques et d'enquêtes qualitatives ;
- que les personnes ne puissent faire l'objet d'investigations plus poussées à partir de la base de données sans leur accord ;
- que toute information facilitant une forme de profilage des risques permettant d'individualiser davantage la couverture maladie soit exclue de la publication ;
- que le tout s'opère sous le contrôle d'un comité analogue au CSS, représentant les différents intérêts en présence : personnes enquêtées, organismes produisant des statistiques, professionnels de santé, chercheurs, État. Ce comité aura pour charge de veiller au respect de chacun des points précédents.

L'ensemble de la démarche vise donc à permettre un véritable accès aux données de santé, tout en préservant les droits et les intérêts des personnes. L'éthique médicale et celle du chercheur en sciences sociales se retrouvent au moins sur un point – qu'on pourrait qualifier tout simplement de moral dans un sens kantien : avant tout, ne pas nuire. En respectant ce principe, la marge de progression est large en faveur d'un droit à une meilleure connaissance des processus sociaux qui encadrent la santé publique.

Références

Beaud, Stéphane, et Florence Weber. 1997. *Guide de l'enquête de terrain: produire et analyser des données ethnographiques*. 2010 éd. Guides Repères. Paris: La Découverte.

Beliard, Aude, et Jean-Sébastien Eideliman. 2008. « Au-delà de la déontologie. Anonymat et confidentialité dans le travail ethnographique ». In *Les politiques de l'enquête. Epreuves ethnographiques*, édité par Alban Bensa et Didier Fassin, 123 42. Recherches - Bibliothèque de l'Iris. Paris: La Découverte.

Belorgey, Nicolas. 2010. *L'hôpital sous pression: enquête sur le « nouveau management public »*. Textes à l'appui - Enquêtes de terrain. Paris: La Découverte.

Bosa, Bastien. 2008. « A l'épreuve des comités d'éthique. Des codes aux pratiques ». In *Les politiques de l'enquête. Epreuves ethnographiques*, édité par Didier Fassin et Alban Bensa, 205 27. Recherches - Bibliothèque de l'Iris. Paris: La Découverte.

Bruno, Isabelle, et Emmanuel Didier. 2013. *Benchmarking: l'État sous pression statistique*. 1 vol. Paris: Zones.

Desrosières, Alain. 2005. « Décrire l'Etat ou explorer la société: les deux sources de la statistique publique ». Genèses 58: 4 27.

Didier, Emmanuel. « Mettre en responsabilité. "Compstat" et le nouveau management statistique de la Préfecture de police de Paris ». Document de travail.

Gramain, Agnès, et Florence Weber. 2001. « Ethnographie et économétrie: pour une coopération empirique ». Génèses 44: 127 44.

Laurens, S. et Neyrat, F. éd., 2010. *Enquéter : de quel droit ? : Menaces sur l'enquête en sciences sociales*, Bellecombe-en-Bauges: Editions du Croquant.

Passeron, Jean-Claude. 1991. *Le raisonnement sociologique*. 2006 éd. Paris: Nathan.

Weber, Florence. 2009. *Manuel de l'ethnographe*. Paris: PUF.

Weber, Florence., 2011. La Déontologie Ethnographique À l'Épreuve du Documentaire. *Revue de Synthèse*, 132(3), p.325 349.

Weber, Max. 1917. *Essais sur la théorie de la science*. 1965 éd. Paris: Plon.

Une protection renforcée de la vie privée et une plus grande liberté pour les chercheurs

Jean-Pierre LE GLÉAU
Inspecteur général honoraire de l'Insee

Nicolas Belorgey propose une critique du numéro 2014-2 de la revue Statistique et Société qui, selon lui, « ne prend pas en compte les besoins d'une démarche rigoureuse aujourd'hui en sciences sociales ».

Il considère en effet que la connaissance en sciences sociales doit s'appuyer sur des outils non seulement quantitatifs, mais aussi qualitatifs. Pour cela, il estime qu'il est nécessaire d'allier deux démarches : le raisonnement sur des cas individuels, afin d'obtenir une bonne compréhension des phénomènes au niveau le plus fin possible, et les tests d'hypothèses statistiques, afin de vérifier le degré de généralité de ces éléments intermédiaires.

L'article intitulé « L'accès aux données confidentielles de la statistique publique – De la sensibilité des données économiques à la sensibilité des données de santé » semble particulièrement visé par ce reproche. Son écriture a sans doute été imprécise, ou mal comprise par le lecteur. En effet, les carences supposées de la démarche qui y est décrite (passage par le comité du secret statistique, puis accès aux données sur un centre sécurisé) constituent justement le point fort de celle-ci, en permettant au chercheur d'effectuer des allers-retours entre les données individuelles et les données agrégées. L'incompréhension provient peut-être du fait qu'il est dit que le centre d'accès sécurisé (CASD) héberge des données de la statistique publique. Il faut bien comprendre qu'il s'agit des données individuelles qui ont permis, après traitement, de produire des statistiques anonymes. Mais le chercheur a accès aux données en amont de ces traitements, les plus détaillées possible qui, bien qu'anonymisées, restent confidentielles à cause des possibilités d'identification indirecte. C'est d'ailleurs pour cela, et afin de protéger la vie privée, qu'un protocole spécifique est nécessaire pour que les chercheurs puissent travailler sur ces données individuelles très détaillées. Ce protocole laisse par contre la plus grande liberté au chercheur dans son travail, et lui permet notamment d'effectuer des allers-retours entre les cas individuels et le cadrage statistique. La seule contrainte qui s'impose à lui est de ne sortir de cette « bulle » que des résultats rendant impossible l'identification d'une personne, contrainte qui semble difficilement contestable.

L'auteur de la critique suggère aussi que les données administratives déjà disponibles doivent pouvoir être enrichies par des informations provenant d'enquêtes réalisées par le chercheur. C'est effectivement une démarche très féconde. Mais celle-ci est tout à fait réalisable dans le cadre du protocole décrit dans l'article. Des appariements de données provenant de diverses sources (administratives ou personnelles) peuvent être réalisés dans l'enceinte du CASD. Ce n'est même que depuis que ce centre d'accès sécurisé a été mis sur pied que ces appariements sont devenus possibles. Certes, les démarches nécessaires pour réaliser de tels appariements sont souvent complexes, lourdes et longues à mettre en œuvre. Mais cette complexité n'est pas à imputer au protocole du CASD, mais tout simplement… au texte de la loi elle-même (notamment la loi informatique et libertés) qui impose des processus très lourds pour les

appariements de fichiers, en particulier ceux contenant le numéro d'inscription au répertoire d'identification des personnes physiques (NIR). Un assouplissement de ces règles figure dans le projet de loi sur la santé qui sera prochainement examiné par le Parlement.

L'auteur propose in fine un autre type de démarche : sélectionner un certain nombre de personnes à partir de données statistiques et compléter les informations sur celles-ci à l'aide d'enquêtes qualitatives. Cette démarche est certainement plus délicate, puisqu'elle suppose de « sortir » de la « bulle » du CASD des données statistiques individuelles pour pouvoir interroger les personnes. Cela est contraire à l'esprit général du processus décrit dans l'article de 2014. On conviendra volontiers que ce plan d'expérimentation, qui consiste à choisir des individus, avec leurs données confidentielles, pour pouvoir les interroger comporte des risques sérieux pour la protection de la vie privée. Il n'est cependant pas formellement impossible, bien que ne s'inscrivant pas dans le processus standard développé pour le CASD. Il suppose l'intervention d'un tiers de confiance et des protocoles complexes, qui ont cependant déjà été mis en œuvre, mais pour l'instant au seul profit de centres de recherche publics.

Comme on le voit, les voies proposées par l'auteur, loin d'être incompatibles avec la démarche « comité du secret – CASD » sont rendues possibles précisément depuis que cette architecture a été mise en place. Il en résulte une protection renforcée de la vie privée et une plus grande liberté pour les chercheurs. Celle-ci est d'ores et déjà appréciée par les chercheurs en sciences économiques ou sociales. Elle pourrait l'être à l'avenir des chercheurs travaillant dans le domaine de la santé.

Quelles données de santé pour quelle recherche ?

Marcel GOLDBERG et Marie ZINS
Épidémiologistes, Inserm-Université de Versailles-Saint-Quentin[1]

Nicolas Belorgey plaide dans son article pour un « véritable accès aux données de santé », permettant de pouvoir utiliser à la fois les données individuelles issues de traitements statistiques enregistrées dans les bases d'origine administrative et des données recueillies directement auprès des personnes. Il argumente que les données d'origine administrative ne sont qu'un sous-produit de l'activité des organismes qui les produisent et sont loin de refléter la réalité des phénomènes ainsi mesurés, malgré leur intérêt évident. On ne peut que souscrire à ce point de vue, de bon sens scientifique évident. Cependant, la vision développée par N. Belorgey appelle quelques commentaires.

Ainsi, il nous semble qu'il y a confusion entre recueil de données directement auprès des personnes et méthodes qualitatives relevant des sciences sociales. Il s'agit en fait de deux aspects distincts, qui concernent d'une part l'obtention de données directement auprès de personnes, et d'autre part les méthodes de recueil de ces données.

En tant qu'épidémiologistes, nous ne pouvons adhérer à l'assimilation de ces deux aspects. L'essentiel de l'activité de recherche de notre groupe consiste en effet dans la conduite de cohortes épidémiologiques, comme Gazel [1] et Constances [2] qui associent données issues des bases administratives et données recueillies directement auprès des personnes, notamment par questionnaires, sans contact personnel avec les participants de nos cohortes, par des méthodes qui ne sont donc pas qualitatives ; données que nous traitons de façon quantitative. Nous considérons nous aussi que les bases de données administratives ne sont pas suffisantes pour répondre à de nombreuses questions de recherche du fait de leurs procédures de constitution qui ne permettent qu'une vision incomplète et partiellement déformée des phénomènes que nous cherchons à analyser. Mais, pour la plupart des recherches épidémiologiques, les limites des bases de données administratives ne sont pas tout à fait de la même nature que pour les sciences sociales, bien que nous soyons nous aussi convaincus que les méthodes qualitatives, isolément ou en association avec des méthodes quantitatives, sont indispensables pour la compréhension de certains phénomènes.

N. Belorgey propose diverses mesures permettant de recueillir sur les mêmes personnes des données pertinentes (d'origine administrative et d'origine personnelle) dans des conditions respectant l'anonymat des personnes lors de la diffusion des résultats des travaux de recherche.

1. Unité mixte de service « Cohortes épidémiologiques en population »

Ces propositions sont générales et sous cette forme, nous surprennent un peu dans la mesure où pour certaines elles correspondent à la situation actuellement en vigueur, et où d'autres contraintes majeures ne sont pas abordées par N. Belorgey. Ainsi, la quasi-impossibilité d'utiliser le NIR[2] (sauf décret en Conseil d'État) qui est, directement ou sous forme cryptée, l'identifiant utilisé dans ces bases ne permet pas d'accéder aux données enregistrées pour une personne dans les bases administratives, même avec son consentement. De plus, s'agissant du SNIIRAM[3], le protocole de cryptage utilisé pour enregistrer les données d'une personne a été conçu pour interdire toute possibilité d'identification directe de la personne concernée, ce qui ne permet pas son utilisation dans les cas où l'enquête commence par l'accès aux données statistiques pour être complétée par des données recueillies auprès des personnes, comme le préconise N. Belorgey.

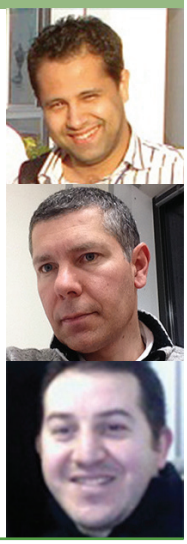
Le premier de ces obstacles devrait être levé si la disposition qui prévoit que la Cnil puisse autoriser l'utilisation du NIR sans passer par le Conseil d'État, prévue par le projet de loi de santé, est retenue par le Parlement qui doit prochainement l'examiner. Il est également prévu dans le même article du projet de loi la mise en place d'un Comité d'expertise pour la recherche, les études ou l'évaluation dans le domaine de la santé pour les demandes d'autorisation, qui se substituerait à l'actuel CCTIRS[4], en en élargissant la composition, notamment pour l'ouvrir à des personnes choisies en raison de leur compétence en sciences sociales.

Références

[1] www.gazel.inserm.fr
[2] www.constances.fr

2. Numéro d'identification au Répertoire national d'identification des personnes physiques (RNIPP)
3. Système d'informations interrégimes de l'assurance-maladie
4. Comité consultatif sur le traitement de l'information en matière de recherche dans le domaine de la

Les trois défis du *Big Data*[1]
Éléments de réflexion



Khalid BENABDESLEM[2]
Christophe BIERNACKI[3]
Mustapha LEBBAH[4]
Groupe « Data mining et apprentissage » de la SFdS

Genèse informatique du Big Data

Le phénomène qu'on appelle aujourd'hui communément le « *Big Data* », ou données massives, appelle une vision globale, c'est-à-dire non limitée aux seuls aspects informatiques par exemple, même si ce phénomène tire essentiellement son origine dans l'accroissement des moyens informatiques et numériques à un coût toujours plus réduit. En effet, le coût de stockage par Mo est passé de 700$ en 1981 à 1$ en 1994 puis à 0.01$ en 2013[5] (prix divisé par 70 000 en une trentaine d'années) tandis que l'on trouve maintenant des disques durs de l'ordre de 8 To à comparer aux 1.02 Go de 1982[6] (capacité multipliée par 8 000 sur la même période) et une vitesse de traitement pour l'ordinateur le plus performant du moment passant d'un gigaFLOPS (le FLOP correspond à *FLoating-point Operations Per Second*) en 1985 à plus de 33 petaFLOPS en 2013[7] (vitesse multipliée par 33 millions).

Il faut bien avoir conscience qu'aucun domaine n'échappe à cette accumulation de données numériques, ce qui justifie pleinement l'intérêt d'un aperçu global. On peut citer une liste bien longue mais qui donne l'ampleur sociétale du phénomène : commerce et affaires (système d'information d'entreprise, banques, transactions commerciales, systèmes de réservation…), gouvernements et organisations (lois, réglementations, standardisations, infrastructures…), loisirs (musique, vidéo, jeux, réseaux sociaux…), sciences fondamentales (astronomie, physique et énergie, génome…), santé (dossier médical, bases de données du système de sécurité sociale…), environnement (climat, développement durable, pollution, alimentation…), humanités et sciences sociales (numérisation du savoir, littérature, histoire, art, architectures, données archéologiques…). Toute la société converge ainsi vers un monde numérique, au point qu'en 2007 plus de 94% de l'information stockée l'était sous forme numérique (les 6% restants sous forme analogique), à comparer à seulement 1% en 1986 (voir la figure 1). En outre, cette

1. Compte rendu de la journée thématique du 13 mars 2015 organisée par la SFdS. Le descriptif de la journée est disponible ici : http://www.sfds.asso.fr/393-Big-Data. On trouvera aussi un lien vers les présentations de cette journée sous forme de pdf et de vidéos (la journée avait été retransmise en direct sur le web).
2. Université Lyon 1, CNRS UMR 5205 LIRIS
3. Université Lille 1, CNRS UMR 8524 Painlevé, Inria
4. Université Paris 13, CNRS UMR 7030 LIPN
5. http://www.capital.fr/enquetes/documents/la-folle-evolution-du-stockage-informatique-953110
6. http://fr.wikipedia.org/wiki/Disque_dur#.C3.89volution_en_termes_de_prix_ou_de_capacit.C3.A9
7. http://fr.wikipedia.org/wiki/FLOPS

quantité d'information stockée dépasse maintenant les 280 Eo (exaoctet), contre 0.02 Eo en 1986 (14 000 fois plus).

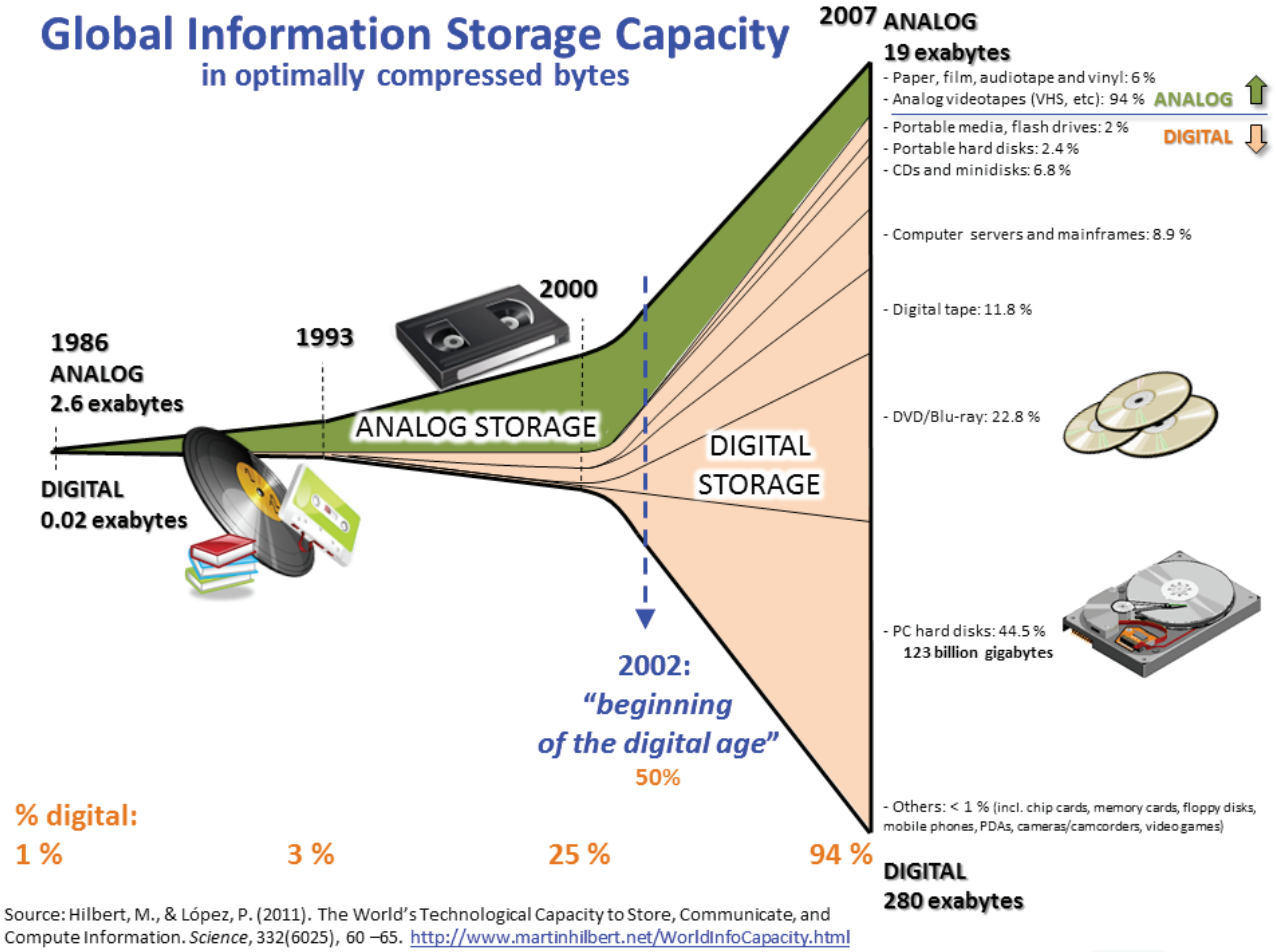


Figure 1. Evolution de la capacité de stockage numérique

Mais cette nouvelle société du « tout numérique » est-elle suffisamment mature pour s'acclimater en profondeur ? Cette avalanche de données pose en effet des grands défis, incontournables, qui ne sont pas totalement résolus à ce jour. Ils sont de trois types : (i) Le stockage et la préservation à long terme et sa pérennisation, (ii) la gestion et l'analyse adéquate en un temps raisonnable et (iii) l'impact sociétal et économique. Ils englobent, sous une autre typologie, les habituels « 5V » qui caractérisent généralement le *Big Data* : Volume, Vitesse, Variété, Véracité et Valeur. La journée thématique du 13 mars 2015, intitulée « *Big Data*: une vision globale: Gestion, Analyse, Éthique et Logiciels », a permis d'aborder l'ensemble de ces aspects par des acteurs spécialisés. Cette journée a permis de rendre compte que ce mouvement de « *Big Data* », qui est très profond, ne peut pas se limiter uniquement aux cinq caractéristiques ci-dessus. C'est aussi une opportunité pour un dialogue interdisciplinaire inédit, suscitant de vraies questions éthiques et juridiques.

Trois défis pour le *Big Data*

Défi du stockage

Comme discuté plus amont, les données massives proviennent essentiellement des facilités d'acquisition et de stockage des données. Cependant, le volume est encore appelé à croître de façon très rapide, par exemple en astrophysique avec le projet Gaïa (2013) ou le projet Euclid

(2021) qui prévoit d'atteindre de l'ordre de 50 Go par jour d'acquisition de nouvelles données. De façon connexe, il ne sert à rien de stocker ces informations si la performance des accès (transfert typiquement) n'est pas garantie pour un futur traitement par exemple, ou encore si leur disponibilité n'est pas assurée. La question de leur protection et de leur préservation à long terme, pour les générations futures, est également posée.

Défi de l'analyse

La facilité de stockage massif conduit inévitablement à peu de sélection *a priori* sur les données à acquérir. Cela peut être vu comme une véritable chance, permettant de garder toute latitude sur les futurs usages potentiels et qui ne sont en fait pas toujours totalement définis au moment même de leur acquisition.

En particulier, de nombreuses questions qui étaient considérées comme hors de portée auparavant deviendront accessibles, avec à la clé une plus-value potentielle importante (avantage compétitif par exemple).

Il faut cependant garder à l'esprit que des données plus massives ne sont pas toujours de meilleures données. Cela dépend si elles sont ou non bruitées, et si elles sont représentatives de ce qui est recherché. En sus, lorsque le nombre de variables croît, le nombre de corrélations erronées croît également. La partie analyse devra prendre en considération ces aspects essentiels.

Seront aussi stockées des données hétérogènes (structurées, non-structurées) ou encore des données incomplètes ou incertaines pour lesquelles des traitements spécifiques sont nécessaires. A ce sujet d'ailleurs, des traitements spécifiques sont déjà requis pour les données plus standard, le volume des données posant déjà en lui-même des difficultés théoriques et pratiques inconnues jusqu'alors, même si certaines vieilles méthodes restent efficaces. Ainsi, les simples tests statistiques[8] deviennent inopérants pour de grandes tailles d'échantillons. On peut citer aussi la difficulté d'analyses multidimensionnelles sur des grands ensembles de données, problème parfois appelé « fléau de la dimension ». Au-delà de l'extraction des connaissances se pose aussi leur interprétation, la visualisation étant jusqu'à présent un outil extrêmement puissant mais qui risque de devenir inopérant par effet de saturation graphique tout simplement. En sus, l'analyse en temps réel de flux continus de données émanant de différentes sources pose elle aussi des difficultés spécifiques. Toutes ces questions impliquent la mise au point de nouvelles statistiques pour le *Big Data*, nécessitant de revoir par exemple des calculs de base comme les tests statistiques et les corrélations[9]. Par rebond, de nouveaux profils de statisticiens pour les mettre en œuvre devront être définis.

Ces outils d'analyse méthodologiques ne peuvent bien entendu être dissociés des outils informatiques et d'écosystèmes dédiés au *Big Data* comme NoSQL, Hadoop, MapReduce ou encore Spark.

Défi sociétal et économique

Le phénomène de *Big Data* ne concerne pas que l'informaticien ou le statisticien. La protection de la vie privée, le droit à l'oubli, les droits de propriété, les droits d'exploitation, le coût énergétique du stockage ou du transfert sont autant de questions touchant le plus grand nombre. D'un point de vue économique, la définition du rôle pris par ces données est aussi posée : matière première ? produits dérivés ? ou capital ? Avec à la clé leur valorisation économique. La fin des monopoles durables est également possible, la donnée permettant de contourner les traditionnelles barrières à l'entrée que représentent par exemple la détention exclusive de

7. Raftery, A.E. Bayesian model selection in social research. Sociological methodology 25, 111-164, 1995.
8. Meyer-Schönberger, V. & Cukier, K. Big Data: A Revolution That Will Transform How We Live, Work, and Think. Eamon Dolan, 2013.

ressources technologiques ou encore certaines connaissances autrefois monopolisées par les pouvoirs publics. Certes aujourd'hui l'innovation et la découverte de nouveaux procédés avec l'analyse des données créent pour l'entreprise qui en est l'auteur une situation provisoire de monopole mais cette situation est amenée à disparaître rapidement, au rythme accéléré des innovations concurrentes[10]. D'un point de vue sociétal, le statut des données entre propriété privée, domaine public et objet commercial reste souvent flou. Les communautés auto-régulées et le développement transnational vont aussi bouger les lignes actuelles par la grande fluidité de l'information. La gestion fine des déplacements des individus, leur profilage et leur ciblage, le travail sur des recensements plus que des sondages soulèvent des questions de protection des libertés individuelles. D'un point de vue juridique et fiscal, le rôle des États et de leurs instances officielles de surveillance, rôle exercé en France par la Commission Nationale Informatique et Libertés (CNIL), peut changer avec des pertes de ressources fiscales et une pertinence amoindrie des normes juridiques. Ce ne sont pas nécessairement les principes de la loi Informatique et Libertés qu'il faut remettre en cause mais c'est assurément les outils de la régulation qu'il faut adapter. C'est ainsi dans un état d'esprit pragmatique, ouvert et soucieux d'accompagner l'innovation que s'inscrit dorénavant la CNIL[11]. D'un point de vue sécuritaire, on ne peut éviter de penser à une société de surveillance ou de contrôle, symbolisée par Prism et la NSA.

Vers une science des données

La disponibilité de très grandes masses de données et les capacités computationnelles de les traiter de manière efficace sont en train de modifier la manière dont nous faisons de la science. L'informaticien et le statisticien ont pris conscience de l'émergence d'une science des données (*data science*), caractérisée par une collecte massive et variée de données associée à des méthodes de traitements pour en extraire des connaissances nouvelles. A la clé de ce changement de paradigme scientifique se profile aussi un bouleversement de l'enseignement, pour former non seulement les futurs acteurs de cette discipline mais également les citoyens de ce nouveau monde connecté.

Le groupe « DMA » de la SFdS

Le groupe Data Mining et Apprentissage (DMA) de la SFdS vise à tisser des liens étroits entre l'informatique et la statistique, à favoriser les interactions entre théorie, méthodologie et applications, à travailler au renforcement des collaborations entre les membres de la SFdS et ceux d'autres associations savantes connexes. A ce titre, il regroupe des membres de ces différents horizons : informaticiens, statisticiens, praticiens et théoriciens.

10. http://www.creg.ac-versailles.fr/IMG/pdf/La_concurrence_imparfaite.pdf
11. http://www.cnil.fr/linstitution/actualite/article/article/enjeux-2015-2-la-protection-des-donnees-cle-de-voute-de-linnovation/

Big data - La révolution des données est en marche
Recension de l'ouvrage

Antoine CHAMBAZ[1], Isabelle DROUET[2]
Statisticien et philosophe des sciences,
Universités Paris-Ouest Nanterre et Paris-Sorbonne

L'ouvrage dont nous proposons une recension est destiné au grand public et porte sur les données massives, ou ***Big data*** selon un anglicisme bien ancré que nous adoptons ici. Ses auteurs, Victor Mayer-Schönberger et Kenneth Cukier, sont respectivement universitaire spécialiste d'internet et des *Big data* et éditeur associé au magazine *The Economist*. Enthousiastes, ils adoptent un ton assez radical et prophétique, quoique peut-être de façon plus prononcée dans la version française. Un grand nombre de recensions antérieures sont disponibles dans des publications d'horizons divers.

Que sont les *Big data*? La variété des acceptions de l'expression est bien illustrée par la liste de quarante définitions proposées par autant de personnalités sur le blog datascience@Berkeley. La tâche de définition est d'autant plus ardue que la notion de *Big data* continue d'évoluer. S'ils ne tranchent pas définitivement la question, les auteurs proposent néanmoins des éléments de réponse. Les *Big data* peuvent être massives, mais aussi-pêle-mêle-ouvertes, exhaustives, recueillies à l'insu des sujets, désordonnées. Elles peuvent gagner à être croisées, elles sont tournées vers la prédiction. Elles ont de la valeur, et celle-ci peut se révéler bien après leur collecte. Aucune de ces caractéristiques n'est nécessaire, aucune d'elles n'est suffisante.

Toutefois, les auteurs s'intéressent moins à ce que sont les *Big data* qu'à ce qu'elles font. Les Big data permettent d'accomplir à très grande échelle des choses qui ne peuvent pas l'être à une échelle plus petite. Ce faisant, elles créent de nouvelles intuitions, de nouvelles formes de valeur, elles modifient les marchés, et les relations entre citoyens et gouvernements. Une part importante de l'ouvrage est consacrée aux transformations que les *Big data* nourrissent et au mouvement de fond de leur production effrénée, que les auteurs appellent « datafication ». Nous pourrions traduire ce terme par mise en données du réel (ou mise au pas ? Car, n'est-ce pas une vision asséchante de la réalité qui est à l'œuvre quand nous croyons capturer toute sa richesse dans une suite de 0 et de 1 ?). La construction de l'ouvrage révèle discrètement la centralité du concept de « datafication ». Il donne son titre au cinquième des dix chapitres, et ce titre est le seul néologisme quand les neuf autres sont des mots courants.

L'ouvrage est émaillé de nombreux exemples qui viennent à l'appui des thèses des auteurs. Certains sont bien connus. C'est le cas en particulier du projet *Google Translate*. Schématiquement, ce projet abandonne l'approche linguistique traditionnelle de la traduction

1. Laboratoire Modélisation aléatoire Modal'X, Université Paris Ouest Nanterre
2. Equipe de recherche SND Sciences Normes Décision, Université Paris Sorbonne

et adopte une stratégie fondée sur l'alignement phrase à phrase de textes dont nous savons qu'ils sont traductions l'un de l'autre. Le procédé est efficace grâce à la masse de traductions disponibles. Il permet notoirement de traduire d'une langue A à une langue B en utilisant de manière essentielle une langue C comme pivot. Nul besoin de disposer de traductions de A vers B ! D'autres exemples sont également remarquables. A nos yeux, l'élaboration au long cours de la géographie physique des océans entreprise par le commandant Maury, un précurseur du XIXe siècle, et la recherche raisonnée à New-York, par Flowers et son équipe, des logements sur découpés, très exposés au risque d'incendie notamment, comptent parmi les plus marquants. Le second est à double tranchant. Certes, Flowers et ses magiciens des nombres ont augmenté l'efficacité du service de prévention des incendies, et ce faisant ils ont certainement sauvé de nombreuses vies. Cependant, ils ne s'attaquent pas à la précarité du logement à la racine, c'est-à-dire à ses causes, et donc cette précarité est vouée à se perpétuer. Pour la combattre, il est impossible de faire l'économie d'une analyse qualitative, géographique et sociologique, de la ville. Certains exemples se révèlent fragiles dès le plan opérationnel, et plus particulièrement le projet *Google Flu Trends*. Si ce projet a permis de prédire l'évolution épidémique aux Etats-Unis de la grippe H1N1 plus efficacement que les statistiques gouvernementales en 2009, sa fiabilité s'est avérée décevante à l'épreuve du temps.

La thèse principale de l'ouvrage est la suivante: une révolution est en marche qui, selon le sous-titre, bouleversera la façon dont nous vivons, travaillons et pensons. Pour le dire dans les termes de Kuhn (Kuhn, 1962), nous assistons à un changement brutal de paradigme. Selon les auteurs, les caractéristiques principales de ce nouveau paradigme sont (*i*) un glissement d'intérêt du *pourquoi* vers le *quoi*, et de la causalité vers la corrélation (Chambaz et al., 2014), (*ii*) l'éventuelle exhaustivité des données, à laquelle ils associent l'expression N = *tous*, (*iii*) la possibilité désormais pleinement envisageable de laisser parler les données, comme si elles agissaient de manière autonome.

Cette analyse est discutable. Comme le souligne par exemple Strasser (Strasser, 2012), dès la Renaissance les naturalistes ont été submergés par un déluge de données et ont dû développer des techniques pour les classer et les étudier. Le XIXe siècle a lui aussi connu son *avalanche de nombres*, pour reprendre l'expression de Hacking (Hacking, 1982). Par ailleurs, il est un peu vain de comparer des quantités de données au-dessus du fossé béant qui sépare les mondes analogique et digital. En outre, la prétention à l'exhaustivité est très souvent illusoire, si ce n'est toujours, quoi qu'en disent les descriptions généralement emphatiques des bases de données massives. Leonelli démontre ainsi (Leonelli, 2014) que l'utilisation des *Big data* en biologie des organismes modèles n'a pas les caractéristiques mises en évidence par les auteurs. Elle souligne en particulier que les bases constituées sont loin de contenir toutes les données disponibles et souffrent par ailleurs nécessairement d'un biais de représentation. Finalement, indépendamment de l'exhaustivité, lorsque les masses de données sont trop conséquentes, leur manipulation requiert le concours d'artifices techniques tels que des sondages ou des projections. Nous sommes loin, à l'évidence, d'être en mesure de les laisser parler. De manière générale, il nous semble que l'analyse proposée par Mayer-Schönberger et Cukier est largement dépendante de la nature des exemples abordés, marqués par un objectif pratique plutôt qu'épistémique.

Nous aimerions maintenant revenir sur l'importance que les auteurs accordent à N = *tous*. Elle s'explique facilement par la croyance que lorsque nous avons toutes les informations relatives à un phénomène, il devient possible de faire des prédictions exactes. Ce faisant, elle témoigne du présupposé qu'il n'y a rien au-delà des données elles-mêmes, une vision de la réalité très *humienne*, dirait la philosophe, et très *pearsonienne*, dirait le statisticien. Cette vision nous semble aller à rebours du mouvement d'érosion du déterminisme, identifié par Hacking (Hacking, 1990), qui au cours des 18è et 19è siècles a peu à peu placé le hasard au centre de l'échiquier scientifique.

Selon cette vision, il n'y en particulier pas de loi immanente qui présiderait à la production des données. Par conséquent, l'objectif n'est pas de mettre au jour des traits d'une loi immanente mais seulement d'élaborer des descriptions factuelles des données disponibles et de faire des prédictions.

Pour bien mettre en lumière les ressorts de cette position, considérons l'exemple de la prédiction temporelle, c'est-à-dire de l'anticipation du futur sur la base du passé. Elle consiste à s'appuyer sur des descriptions factuelles de données recueillies à des instants successifs (nous parlons alors de données temporelles), puis à déterminer les tendances qui s'en dégagent et, enfin, à les prolonger. Dans la théorie statistique de l'estimation séquentielle, c'est justement l'existence d'une loi immanente qui justifie l'opération de prédiction, éventuellement au prix d'hypothèses dont la formulation même engage une référence à cette loi. Dans cette théorie, différents prédicteurs peuvent être mis en concurrence par comparaison des risques qui leur sont associés, le risque d'un prédicteur étant l'écart moyen, sous la loi immanente, entre la valeur à prédire et sa prédiction. Parce qu'elle suggère l'absence de loi immanente, l'importance accordée par les auteurs à N = *tous* les met ainsi en porte-à-faux avec la théorie statistique de l'estimation. Elle les rapproche, en revanche, de la théorie de la prédiction de suites individuelles, où les observations sont le produit d'un certain mécanisme inconnu, non spécifié, pouvant être tout autant déterministe que stochastique, ou même dynamique et antagoniste. Dans ce cadre, ce n'est pas le risque qui est la mesure centrale de qualité d'un prédicteur mais le regret, qui consiste en une comparaison entre les gains effectifs générés par le prédicteur et ce qu'auraient pu être les gains si le prédicteur optimal avait été choisi. Notons qu'ici, le hasard est réintroduit de la main des prédicteurs eux-mêmes: chassez le hasard, il revient au galop ! La randomisation se révèle en effet essentielle pour déterminer un prédicteur optimal.

Bien qu'ils soient partisans de laisser parler les données, les auteurs nous préviennent: "nous devons nous garder de trop nous reposer sur les données pour ne pas répéter l'erreur d'Icare qui, subjugué par la prouesse technique du vol, ne l'a pas maîtrisé et s'est abîmé en mer". Loin des considérations épistémiques qui ont été les nôtres, ils situent les limites du recours aux *Big data* sur le seul terrain réglementaire ou légal. Les inquiétudes qu'ils expriment concernent principalement la vie privée et les libertés individuelles. Un mouvement de fond économique et sociétal est assurément en marche. Il nous appartient de le contrôler et, surtout, de nous l'approprier. L'État français a ainsi créé une plateforme ouverte des données publiques françaises[3] et appelle à y contribuer librement. Nous faisons collectivement face à un enjeu fondamental d'éducation et de sensibilisation de tous les citoyens, jeunes et moins jeunes. Ensemble, domptons les *Big data*, ne nous laissons pas dévorer par elles. Nous concluons la recension sur cette note résolument optimiste.

Références

Chambaz, I. Drouet, and J-C. Thalabard. *Causality, a trialogue*. Journal of Causal Inference, 2 (2): 201-241, 2014.

Hacking. *Biopower and the avalanche of printed numbers*. Humanities in Society, 5 (3-4): 279-295, 1982.

Hacking. *The taming of chance*, volume 17. Cambridge University Press, Cambridge, 1990.

T. S. Kuhn. La structure des révolutions scientifiques. 1962.

S. Leonelli. *What difference does quantity make? On the epistemology of Big Data in biology. Big Data & Society*, 1 (1): 1-11, 2014.

J. Strasser. *Data-driven sciences: From wonder cabinets to electronic databases. Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, 43 (1): 85-87, 2012.

3. https://www.data.gouv.fr/fr/

MANIFESTE[1]
Pour l'intégration des activités de lien social dans la nomenclature des catégories socio-professionnelles

Antoine HOULOU-GARCIA
Auteur du livre Le monde est-il mathématique ?[2]

La nomenclature est une chose bien connue du statisticien : c'est l'outil qui lui permet de ventiler un résultat global en résultats partiels plus précis. Ainsi, toute étude statistique traitant de l'économie va nécessairement faire appel à la nomenclature d'activités française, la fameuse NAF, qui permet de déterminer les contours de l'industrie, du commerce, des services etc. C'est un outil tellement basique et fondamental pour le statisticien que le site internet de l'Insee[3] dédie une page spécifique aux nomenclatures. Mieux encore : quand on clique sur l'onglet « Définitions et méthodes », accessible dès la page d'accueil du site, on arrive sur une page qui ne commence ni par l'item « Définitions » (en 2ème position) ni par l'item « Sources et méthodes » (3ème position) mais bien par l'item « Nomenclatures ». En cliquant sur cet item, on voit apparaître une liste de nomenclatures de type économique, juridique et socioprofessionnel.

Mais la nomenclature n'a historiquement rien à voir avec la statistique, l'économie ni même le droit. Car la nomenclature, c'est d'abord un mot, et un mot à l'étymologie particulièrement éclairante pour nous faire comprendre ce qu'il sous-tend. Initialement, la *nomenclatura*, en latin, c'est le fait de « faire l'appel » à l'armée en appelant par son nom chacun des soldats. Dans le terme nomenclature, on retrouve le latin *nomen* qui désigne le nom et le verbe latin *calo* qui signifie convoquer. Il est d'ailleurs intéressant de noter que le verbe *calo* est à rapprocher du latin *classis* qui désigne autant la classe d'école que l'armée, deux lieux qui ont notamment en commun de faire l'appel[4].

Par la suite, on a appelé *nomenclator* l'esclave qui, toujours aux côtés de son maître, lui soufflait à l'oreille le nom des gens qu'il rencontrait lors notamment des salutations matinales. Le *nomenclator* permettait ainsi aux plus ambitieux de la République romaine de flatter leur électorat en feignant de se souvenir des noms des électeurs potentiels rencontrés sur le forum, et éventuellement de s'enquérir de la santé du petit dernier (si le *nomenclator* avait lui-même une très bonne mémoire !). « L'information, c'est le pouvoir » pourrait donc être la devise d'un bon *nomenclator*, mais aussi d'un bon nomenclaturiste car être capable d'appeler les gens par

1. Note de la rédaction : cette rubrique accueille des propositions ayant trait à l'utilisation sociale des statistiques. Comme tous les articles de la revue, ces textes n'engagent que leurs auteurs. Comme ils ne visent pas principalement à la description ou à la pédagogie, ils ne font pas l'objet d'une révision éditoriale sur le plan technique avant leur publication.
2. Houlou-Garcia A., Le monde est-il mathématique ? Les maths au prisme des sciences humaines, Honoré Champion, collection Essais, février 2015.
3. http://www.insee.fr
4. Notons que le terme calo est apparenté au grec kaleo, qui a la même signification et qui donne le terme ekklesia, désignant l'assemblée des citoyens à Athènes (ce terme donnant par ailleurs lui-même le françaiséglise). Appeler les gens est donc soit une action d'individualisation soit une action de rassemblement.

leur nom ou leur profession, nous le verrons, est un véritable enjeu de société pour qui souhaite la décrire.

Un outil à double tranchant

La nomenclature, c'est d'abord « la simple dénomination des choses » pour reprendre une expression du grand poète et homme politique Léopold Sédar Senghor[5]. Cette dénomination, elle se fait par exemple dans la toponymie : donner un nom unique et reconnu par tous à une rivière, à une montagne, à un village, c'est faire le travail de nomenclature basique qui permet de connaître le paysage géographique, tout comme établir une nomenclature des activités permet d'approcher le paysage économique. Mais, en donnant un nom précis à chaque chose, on enferme les choses en question dans une dénomination figée, risquant aussitôt de s'éloigner avec le temps de la réalité que l'on veut décrire. Ainsi, les fleuves bougent et, exemple célèbre pour l'historien, les fleuves Tigre et Euphrate de la Mésopotamie se rejoignent de nos jours alors qu'ils ne se touchaient guère il y a quelques millénaires.

Pire encore : la nomenclature est un risque pour ceux qu'elle concerne, directement ou indirectement. En effet, nommer un fleuve, puis une montagne puis une route permet à l'ennemi de venir dans vos contrées car, en nommant les choses, vous lui donnez une carte précise pour venir vous inquiéter. C'est le cas, par exemple, en Asie du Sud-Est, dans la zone reculée de la Zomia, zone habitée par des gens ayant fui l'oppression des Etats (impôts, devoirs militaires etc.) pour revenir à des régimes moins contraignants. Dans cette zone montagneuse difficile à explorer même pour les gouvernements des pays sur lesquels cette zone a-étatique se déploie, l'administration chinoise déplore que « une rivière peut, en l'espace de cinquante kilomètres, se voir attribuer cinquante noms, et un campement s'étendant sur un kilomètre et demi, trois désignations. Voilà pour ce qui est du manque de fiabilité de la nomenclature ! »[6].

Néanmoins, si la nomenclature est un danger potentiel dans la connaissance qu'elle offre aux autres, elle est également un moyen de re-connaissance pour qui veut de la visibilité. C'est le cas bien connu pour les nomenclaturistes de l'item 25.62A de la NAF. Cet item est le fruit d'une revendication pour qu'un métier soit reconnu à part entière et ne soit pas absorbé dans une nomenclature plus générale où il serait noyé. Ce métier, c'est celui du décolletage, une technique particulière d'usinage[7] ayant un fort poids économique dans la vallée de l'Arve en Haute-Savoie. Pour se doter d'une visibilité économique et d'une reconnaissance institutionnelle, les acteurs de cette activité ont fait en sorte que leur spécificité technique soit inscrite dans le marbre de la nomenclature. Cela leur permet en particulier de pouvoir connaître et faire connaître le chiffre d'affaire de leur activité, sa valeur ajoutée, ses effectifs salariés etc.

Ainsi, la nomenclature, initialement censée représenter la société, est devenue un enjeu pour cette même société et ce, de deux manières complémentaires. D'une part, nous l'avons vu, faire émerger un item de la nomenclature est un enjeu de reconnaissance collectif ; d'autre part, appartenir à un item de la nomenclature est un enjeu individuel. Un exemple important de cette rétroaction de la nomenclature est la volonté de nombreux employés du secteur privé d'acquérir le statut de « cadre »[8]. Il s'agit d'un enjeu économique mais aussi d'un enjeu social qui touche à la reconnaissance du travail effectué, à la reconnaissance des responsabilités et de la qualité du travail fourni. Fatalement, si le statut de cadre est désiré, celui d'employé est considéré comme une première étape dont il faut se départir. Un corollaire de cela réside dans

5. Senghor, député et ministre de la IVe République française, puis premier Président de la République du Sénégal, disait trouver en Mallarmé son goût pour « les-choses-très-cachées » (Éthiopiques, Seuil, Paris 1956, p. 99), et chez Saint-John Perse le sens de la « simple nomination des choses » (Ibid., p. 158). Transcrire la complexité d'une société réelle difficile à modéliser par la simplicité imaginaire de la nomenclature est tout l'art de passer de Mallarmé à Saint-John Perse.
6. Scott. J., Zomia ou l'art de ne pas être gouverné, traduction de Guilhot N., Joly F. et Ruchet O., Seuil, 2013, p. 21.
7. Il s'agit d'une technique de fabrication de pièces dites de révolution (vis, boulon, axe, etc.) par enlèvement de matière à partir de barres de métal.
8. Cf. Luc Boltanski, « Les Cadres La formation d'un groupe social » 1982. Editions de Minuit, collection « Le sens commun ».

le fait que les employés sont moins bien considérés dans la société que les cadres.

Cela provient du fait que nommer, c'est créer une distinction entre ce que l'on nomme positivement (par exemples les catégories supérieures) et ce qui s'en déduit en négatif (les professions moins supérieures). Cette dénomination énonce implicitement un traitement dissymétrique entre ce que l'on regarde et ce que l'on déduit de ce que l'on a regardé (le complémentaire, pour employer un terme mathématique). Dès lors, créer une nomenclature crée *de fait* une nomenklatura au sens, non plus latin, mais russe, cette fameuse nomenklatura que Mikhaïl Voslensky[9] après d'autres a dénoncée en 1970.

Dans l'URSS, la nomenklatura était un ensemble de listes de « camarades dignes de la confiance du Parti » permettant d'appeler[10] certains camarades à des fonctions de responsabilité, chose très enviée car les responsabilités, en plus de conférer une notoriété[11] à ceux qui les avaient, étaient accompagnées d'avantages matériels. Cette nomenklatura était bien sûr définie par une nomenclature précise, une nomenclature établie au regard de l'origine des personnes définissant des classes en lesquelles on pouvait avoir plus ou moins confiance. Ainsi, aux classes « d'origine sociale saine » (descendants d'ouvriers non propriétaires), on opposait les classes « d'origine sociale douteuse » (salariés de l'ancien régime), les classes « d'origine sociale malsaine » (petite bourgeoisie propriétaire) et la classe des « descendants des ennemis du peuple » (aristocratie).

La question de cet article n'est pas de critiquer ce système mais de mettre en évidence le fait qu'il oppose une classe à toutes les autres de même, *mutatis mutandis*, que la catégorie « Cadres et professions intellectuelles supérieures » s'oppose à toutes les autres en ce sens qu'elle est celle que la nomenclature place intrinsèquement comme celle de la haute société économico-intellectuelle. Il est d'ailleurs intéressant de noter que la nomenklatura soviétique a renversé une nomenclature tsariste créée par Pierre le Grand au début du XVIIIe siècle, connue sous le nom de « table des rangs »[12].

Il se trouve très étonnamment que cette nomenclature, émanant d'un régime aristocratique, permettait, par les études, d'atteindre des rangs de noblesse. En effet, un étudiant, lorsqu'il démarrait son parcours scolaire, se voyait attribuer un rang assez bas. Une fois ses études finies, il obtenait un rang plus élevé puis, suite à son parcours professionnel, pouvait atteindre un véritable rang de noblesse. La force de ce système nomenclaturiste était de promouvoir une « Noblesse de service » qui soit à égalité de valeur avec la « Noblesse de sang ». Ce système était donc, de ce point de vue, en avance sur son temps car il reconnaissait ce que Bourdieu nommera bien plus tard le capital culturel.

Brève analyse historique de la nomenclature des métiers

Et c'est justement sur les analyses de Bourdieu que va reposer la refonte de la nomenclature des catégories socioprofessionnelles en 1982 sous l'égide d'Alain Desrosières et Laurent Thévenot[13]. La nomenclature précédente, due à Jean Porte en 1954, était basée sur un regard socio-économique de la société tournant essentiellement autour de l'opposition entre patrons et salariés. Cette nomenclature, dite des catégories socio-professionnelles, avait pour objectif de « classer les individus selon leur situation professionnelle en tenant compte de plusieurs critères : métier proprement dit, activité économique, qualification, position hiérarchique et

9. Nomenklatura : les privilégiés en URSS. Édition originale publiée en 1970, traduction française en 1980.
10. Rappelons que le sens premier de la nomenclature est bien d'appeler les gens.
11. Au sens de « devenir un notable ».
12. Табель о рангах en russe.
13. Une analyse irremplaçable des motivations de cette refonte est présentée par Alain Desrosières lui-même dans son ouvrage posthume préfacé par Emmanuel Didier, Prouver et gouverner - Une analyse politique des statistiques publiques, La Découverte, Paris, 2014. Nous tenterons néanmoins ici de proposer une analyse personnelle de cette question.

statut »[14]. Ainsi, l'agriculture est mise à part car étant considérée comme faisant partie des métiers hérités du passé ; les patrons industriels et commerciaux sont regroupés face à aux employés et aux ouvriers ; les professions libérales et cadres supérieurs sont mis en regard des cadres moyens.

0. Agriculteurs exploitants
1. Salariés de l'agriculture
2. Patrons de l'industrie et du commerce
3. Professions libérales et cadres supérieurs
4. Cadres moyens
5. Employés
6. Ouvriers
7. Personnels de services
8. Autres catégories

Tableau 1. La CSP de Jean Porte (1954)

Cette nomenclature n'est pas un pur produit de l'Insee alors naissant[15] car comme Alain Desrosières et Laurent Thévenot le résumaient : « Le taxinomiste enregistre l'état de ces luttes avec des déformations qui tiennent à la position qu'il occupe »[16]. L'état des luttes mentionné par les deux auteurs n'est pas une métaphore mais bien une référence à la question salariale mise en lumière dès 1936, suite à l'accession au pouvoir du Front populaire, et prolongée jusqu'en 1945, au sortir de la guerre. En effet, c'est dans cette période que vont naître les conventions collectives de branches et la nécessité de créer des grilles salariales en fonction notamment des métiers et des niveaux de qualification. Parallèlement naîtra le statut de fonctionnaire en 1941, sous Vichy, dissous puis remplacé par un statut général des fonctionnaires en 1946 avec l'établissement des catégories hiérarchiques A, B, C et D.

Ainsi, quand Jean Porte établit sa nomenclature en 1954, elle reflète réellement les négociations et les luttes récentes sur la catégorisation des emplois. Dès lors, la nomenclature est une photographie de l'imaginaire social, cristallisé autour de l'opposition économique entre les agents. Car, s'il est vrai qu'elle intègre le niveau de qualification, il est abordé en terme de hauteur de revenu : les patrons gagnent plus que les cadres supérieurs qui gagnent eux-mêmes plus que les cadres moyens etc. Si cette vision purement pécuniaire est volontairement grossière, je me permets de la mettre en lumière car elle explique la réalité désirée dans l'inconscient collectif : gagner de l'argent.

Mais la nomenclature des CSP va se confronter à un problème de plus en plus fort : elle ne représente pas les « gens de culture », personnes dont les revenus ne sont pas nécessairement très élevés. Les gens de culture, ce sont notamment les enseignants, les artistes et les intellectuels. C'est ce dont se rendent compte Alain Desrosières et Laurent Thévenot à partir des analyses de Pierre Bourdieu qui avait mis en exergue l'opposition complémentaire entre capital économique et capital culturel[17]. Il est à noter qu'Alain Desrosières et Laurent Thévenot sont deux fonctionnaires (administrateurs de l'Insee) et qu'ils ne se sentent pas particulièrement moins haut placés dans la société que des chefs d'entreprise car eux, tout comme Bourdieu qui est un penseur, considèrent que leur capital culturel a bien plus de valeur que le capital

14. http://insee.fr/fr/methodes/default.asp?page=definitions/nom-categories-socio-profes.htm.
15. L'Insee a été créé en 1946, soit huit ans avant l'établissement de cette nomenclature.
16. Desrosières A. & Thévenot L., « Les mots et les chiffres : les nomenclatures socioprofessionnelles », *Économie et Statistique* n° 110, avril 1979, p. 49-65.
17. Pour la réflexion sociologique d'Alain Desrosières, cf. notamment Armatte M., « Introduction aux travaux d'Alain Desrosières : histoire et sociologie de la quantification », *Statistique et société* Vol. 2, n° 3, novembre 2014.

économique d'un patron.

En voulant mettre en valeur le capital culturel, Pierre Bourdieu, en schématisant volontairement quitte à tomber dans une analyse un peu rapide, veut montrer que par l'intellect il a pu quitter le milieu social de ses parents et, pour le dire ainsi, faire partie du « gratin ». De même, Alain Desrosières et Laurent Thévenot font partie de cette bourgeoisie acquise par l'intellect et se situent certes dans la classe dominante de la société, mais dans la sous-classe dominée de la classe dominante : en effet, la classe dominante de la classe dominante est la grande bourgeoisie économique tandis que la bourgeoisie intellectuelle n'a pas la même reconnaissance sociale.

C'est dans ce contexte et avec cette idée derrière la tête que va émerger en 1982 la nomenclature des « professions et catégories socioprofessionnelles » ou PCS. La nomenclature PCS devrait donc s'appeler PCSP puisque la précédente, « catégories socioprofessionnelles » était abrégée en CSP. Le plus intéressant est néanmoins de voir l'apparition du terme « professions », qui est due au fait que l'on oppose les professions d'une sphère que l'on peut qualifier grosso modo de libérale (« Artisans, commerçants et chefs d'entreprise ») à des professions majoritairement de la sphère publique intellectuelle (« Cadres et professions intellectuelles supérieures »). Notons à propos de cette dernière qu'elle regroupe en son sein une sous-classe intitulée « Cadres de la fonction publique, professions intellectuelles et artistiques », sous-classe qui montre bien le poids de l'intellect dans la refonte de la nomenclature qu'ils proposent alors.

1. Agriculteurs exploitants
2. Artisans, commerçants et chefs d'entreprise
3. Cadres et professions intellectuelles supérieures
4. Professions Intermédiaires
5. Employés
6. Ouvriers

Tableau 2. La PCS d'Alain Desrosières et Laurent Thévenot (1982)

Pour le reste, la création de la catégorie « Professions intermédiaires » est une manière de regrouper les cadres B et assimilés du public comme du privé (on est dans une approche diplôme). Quant aux autres professions, elles ne semblent pas avoir retenu particulièrement l'attention des deux nomenclaturistes qui laissent telles quelles les catégories d'employés et d'ouvriers, classes généralement peu diplômées. Ainsi, comme nous le disions dans la première partie de cet article, le regard porté sur le capital culturel a créé, en complémentaire, un regard dissymétrique sur les gens dans l'ombre de ce même capital culturel : seuls les intellectuels ont « bénéficié » de cette refonte. Qui sont ces gens de l'ombre ?

Pour une reconnaissance des activités de lien social dans la nomenclature

La nomenclature évolue à nouveau en 2003 mais repose toujours sur la même grille de lecture associant le capital économique et le capital culturel, l'énumération des six grandes classes restant inchangée. Comme nous le disions plus haut, la nomenclature est un outil figé par nature tandis que la société est en perpétuelle évolution. Évolution dans sa réalité, dans la représentation qu'elle a d'elle-même, mais aussi dans les rapports de force et les tensions (positives ou négatives) qui l'agitent et l'animent. Notre société a évolué dans les rapports entre les individus qui la composent avec notamment la montée en puissance des seniors qui ont été le moteur du développement des services à la personne.

Autre phénomène d'importance : l'abandon par la République[18] de nombreuses zones, au premier rang desquelles les « cités » des banlieues et les fameuses zones périurbaines. Dans ces zones où les populations sont particulièrement fragiles (financièrement, socialement, scolairement), de nombreuses activités se sont développées. Il n'est pas question ici d'une stigmatisation évoquant une quelconque économie souterraine mais bien au contraire de mettre en lumière les formidables énergies développées par les milieux associatifs et bénévoles. Or il se trouve que dans la nomenclature PCS 2003, aucune mention explicite n'est faite du milieu associatif qui, pourtant, représente une valeur ajoutée considérable, valeur ajoutée difficile à quantifier en termes économiques mais valeur ajoutée humaine, culturelle, éducative et sociale évidente. De même, ce que l'on regroupe sous la bannière « Économie sociale et solidaire »[19] n'est pas approchable directement par la PCS.

Aujourd'hui, dans notre société de plus en plus crispée, dans notre lien social de plus en plus ténu, dans notre pays où les individus s'opposent de plus en plus, il me semble important, pour ne pas dire essentiel à la pérennité de notre vivre-ensemble, de reconnaître la valeur des activités de lien social. Nous parlons bien ici d'activités et non forcément de métiers ni même d'activités rémunérées (ce que l'on nomme « travail »). Ainsi, un retraité pourrait à juste titre être considéré comme ayant une activité de lien social s'il fait partie d'une maison de quartier ou d'une activité bénévole culturelle.

Car il me semble également important, à travers cela, de montrer que les retraités ne sont pas des gens mis au placard de la nomenclature et de la société mais bien des acteurs potentiels de cette société au même titre que tous les autres membres qui la composent. Les retraités doivent d'ailleurs être considérés comme de véritables acteurs dans le cas, de plus en plus développé, des chambres qu'ils louent à des étudiants, dans le cadre également de ces immeubles, gérés par des groupes paritaires de protection sociale, dans lesquels cohabitent étudiants et personnes âgées : certes les étudiants ont une activité de lien social envers les personnes âgées avec qui elles prennent le temps de discuter, qu'elles aident pour certaines tâches etc. Mais il serait faux de considérer que ces personnes âgées n'ont pas, symétriquement, une action de lien social.

Ainsi, en parcourant finement la nomenclature PCS 2003, il me semble que l'on pourrait rassembler sous la bannière « Activités de lien social » les items suivants :

313a	Aides familiaux non salariés de professions libérales effectuant un travail administratif
335a	Personnes exerçant un mandat politique ou syndical
422d	Conseillers principaux d'éducation
422e	Surveillants et aides-éducateurs des établissements d'enseignement
434c	Conseillers en économie sociale familiale
435a	Directeurs de centres socioculturels et de loisirs
435b	Animateurs socioculturels et de loisirs
563a	Assistantes maternelles, gardiennes d'enfants, familles d'accueil
563b	Aides à domicile, aides ménagères, travailleuses familiales
564a	Concierges, gardiens d'immeubles
564b	Employés des services divers

Tableau 3. Première liste des activités de lien social

18. Abandon au moins ressenti si l'on considère qu'il n'est pas toujours réel (mais cette question n'est pas l'objet du présent article).
19. Pour une description précise de ce secteur, cf. Bisault L. & Deroyon J., « L'économie sociale, des principes communs et beaucoup de diversité », Insee Première n° 1522, novembre 2014.

Justifions brièvement la présence dans cette liste des mandats politiques et syndicaux : ce sont des activités qui créent une valeur ajoutée sociale immatérielle particulièrement précieuse quand elles sont bien faites car elles permettent de représenter le peuple et ses aspirations, collectives et individuelles ; ce sont des activités de gouvernance de notre société et d'écoute et elles ont donc toute leur place parmi les activités de lien social.[20]

Ces items issus directement de la PCS pourraient être complétés, comme nous l'avons évoqué, par les activités associatives bénévoles ou rémunérées dans les domaines culturels, politiques, sociaux et éducatifs ; les fondations pourraient d'ailleurs y être intégrées au même titre. S'y ajouteraient les activités de microcrédit et de développement local. Il faudrait réfléchir plus avant à la question des auberges de jeunesse qui sont des lieux de sociabilisation.

Dans un deuxième cercle (peut-être plus discutable au cas par cas), on pourrait également intégrer ces items de la PCS :

423b	Formateurs et animateurs de formation continue
424a	Moniteurs et éducateurs sportifs, sportifs professionnels
434a	Cadres de l'intervention socio-éducative
434b	Assistants de service social
434d	Éducateurs spécialisés
434e	Moniteurs éducateurs
434f	Éducateurs techniques spécialisés, moniteurs d'atelier
434g	Éducateurs de jeunes enfants
441a	Clergé séculier
441b	Clergé régulier
563c	Employés de maison et personnels de ménage chez des particuliers

Tableau 4. Deuxième liste des activités de lien social

Dans cette seconde liste, j'ai souhaité faire apparaître des activités de resociabilisation (éducateurs, assistants sociaux) qui complètent certaines activités de la première liste (en particulier les aides-éducateurs des établissements d'enseignement). Également, et sous un gros point d'interrogation, la présence du clergé au sens général du terme : il ne s'agit bien sûr pas que du seul clergé catholique mais des clergés de toutes les religions. À titre purement personnel, je crois que leur utilité sociale peut être réelle pour nombre de croyants et, bien qu'étant moi-même parfaitement laïc, je pense qu'une nomenclature doit représenter la société telle qu'elle est, ce qui signifie dans ce cas qu'on ne peut passer sous silence la religion et son institution dans la mesure où elle prend une place importante dans la vie de nombreuses personnes.

Notons enfin qu'en ne comptant que les items de la PCS cités dans les tableaux précédents (sans, donc, les retraités notamment), cette nomenclature des activités de lien social regrouperait environ 2 millions de personnes dont environ 80 % de femmes (d'après les données du Recensement de la population).

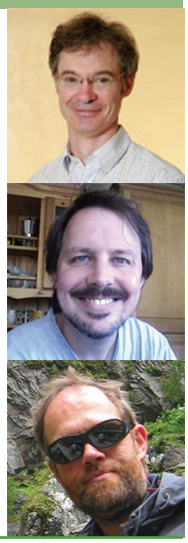
20. La nomenclature ne regarde pas le bon fonctionnement des choses mais se base uniquement sur leur état idéal ; l'argument de la classe politique éloignée des réalités des citoyens, s'il est certes pertinent dans une réflexion sociopolitique, ne l'est donc pas dans le cadre d'une réflexion nomenclaturiste.

Conclusion

Une nomenclature parfaite pourrait sans doute décrire de façon neutre une société entière dans sa complexité. Mais une nomenclature humaine est faite en mettant en lumière certains angles plutôt que d'autres. Comme nous l'avons vu, la première nomenclature avait digéré et intégré les luttes de classes et les négociations salariales ; par la suite, l'analyse « bourdieusienne » du capital culturel y a été incorporée. Désormais, il semble que le poids réel et imaginaire dans notre société du lien social frappe à la porte de la nomenclature. Il ne s'agit pas, à travers cela, de simplement reconnaître l'existence du lien social comme activité, mais également de donner une reconnaissance à ceux qui s'y investissent. Cette reconnaissance institutionnelle leur donnera à terme une reconnaissance professionnelle car on pourra identifier les compétences qu'ils y auront développées.

Adjoindre à la nomenclature existante les activités de lien social n'est pas qu'un geste de reconnaissance d'une réalité, c'est également un geste citoyen. Car les statistiques, nous le savons bien, rétroagissent sur les personnes qui sont elles-mêmes l'objet de ces statistiques. La statistique, en ce sens, n'est pas un objet éthéré coupé de la société : c'est bien au contraire un objet politique au sens étymologique d'acteur de la cité (*polis* en grec). Il est donc important que la statistique soit un acteur et un moteur de notre société. La nomenclature, en particulier, dans son rôle de dénomination et de qualification, est un élément clé de la représentation qu'ont les individus d'eux-mêmes et de leurs concitoyens ; il est donc essentiel qu'elle soit un des moteurs du progrès de notre société. À mon sens, reconnaître dans la nomenclature les activités de lien social serait donc un élément de progrès social.

La recherche reproductible : une communication scientifique explicite



Christophe POUZAT[1], Andrew DAVISON[2], Konrad HINSEN[3]

Chargés de recherche du CNRS en statistique, neurosciences computationnelles, et physio-chimie computationnelle

Dans les articles de recherche ordinaires, tout n'est pas écrit, loin de là : des connaissances sont présupposées, des détails techniques sont omis. Depuis quelques années, des chercheurs ont entrepris de publier des articles qui, en plus du contenu scientifique classique, contiennent toute l'information nécessaire à la reproduction de celui-ci, une fois les données acquises. Des logiciels spécifiques ont été mis au point pour rendre aisée la production de tels articles. Les chercheurs ont intérêt à en profiter, car les revues scientifiques et les agences de financement de la recherche leur demandent de plus en plus d'adopter ce genre de pratiques.

La locution « recherche reproductible » apparaît de plus en plus fréquemment dans des articles scientifiques, des forums ou des blogs voire dans les préoccupations de grandes agences de recherche scientifique comme le National Institute of Health -NIH- (Collins, 2014). Ce phénomène a probablement de quoi surprendre auteurs et lecteurs de littérature scientifique pour qui le qualificatif de « scientifique » entraîne, implicitement au moins, la notion de reproductibilité. Tout lecteur, prenant un peu de recul, va néanmoins très vite réaliser que ce que Michael Polanyi (1998) désignait par « connaissance tacite » joue un rôle – nécessairement – considérable dans la vie quotidienne du scientifique, comme de toute personne d'ailleurs. En statistique, nous pourrions mettre les « conditions de régularité » des théorèmes que nous employons dans la catégorie des « connaissances tacites » ; en substance, ces dernières nous permettent de communiquer de façon concise avec nos collègues ; elles nous fournissent une certaine économie de pensée. La concision qu'elles apportent devient par contre un handicap lorsque nous devons communiquer avec des scientifiques d'un autre domaine, mais aussi avec des non scientifiques : deux situations fréquentes pour les statisticiens. Des difficultés peuvent apparaître également, en interne à un domaine de recherche, lorsque nous essayons d'accéder directement à l'« ancienne » littérature le concernant, puisque dans ce cas, la connaissance tacite peut avoir dérivé au cours du temps. Enfin, avec le recours de plus en plus fréquent à des moyens informatiques importants dans tous les aspects de la recherche, les résultats scientifiques dépendent souvent d'un grand nombre de détails techniques d'un protocole de calcul, qui restent également dans le domaine de la connaissance tacite parce que considérés comme détails techniques (Collins, 2014).

1. MAP5 Université Paris-Descartes et CNRS UMR 8145 , 45, rue des Saints-Pères 75006 Paris , christophe.pouzat@parisdescartes.fr; b.; c.
2. Unité de Neurosciences, Information et Complexité (UNIC; FRE3693 CNRS) 1, avenue de la Terrasse 91198 Gif sur Yvette, andrew.davison@unic.cnrs-gif.fr
3. Centre de Biophysique Moléculaire (UPR4301 CNRS) Rue Charles Sadron, 45071 Orléans Cédex 2, konrad.hinsen@cnrs-orleans.fr

L'objectif

La « recherche reproductible » peut être vue comme une méthode de réduction de l'implicite dans une partie de notre communication. Elle va résulter en un « document dynamique » ou « article actif » (*active paper*) (Hinsen 2014), c'est-à-dire un document qui, en plus de l'article scientifique classique, comportera *toute l'information requise à la reproduction de celui-ci, une fois les données acquises*. Dans la pratique, ce qui est donc entendu par « reproduction » est tout ce qui vient après la collecte des données ; mais comme l'approche requiert un *accès libre* à celles-ci, elles deviennent critiquables et comparables, constituant ainsi un maillon vers une reproductibilité de l'ensemble du processus, qui prend toute son importance dans une période où la production de connaissance repose de plus en plus sur l'utilisation de bases et banques de données collectées par des tiers et rendues accessibles. Plus explicitement, un document dynamique va donner accès à son lecteur, d'une part, à l'ensemble des données brutes sur lesquelles reposent les résultats présentés, d'autre part, à l'ensemble des codes sources développés spécifiquement pour analyser les données et à une description de nature algorithmique de la façon dont les « codes » ont été appliqués aux données. Tout lecteur, s'il le souhaite, pourra alors régénérer l'ensemble des figures et des tables contenues dans l'article, sous réserve qu'il dispose du même environnement logiciel que les auteurs de l'article[4].

L'intérêt

Indépendamment de la justification philosophique qui met l'accent sur la plus grande adéquation entre un idéal scientifique[5] et une pratique quotidienne, il y a d'excellentes raisons, plus banales, pour adopter une pratique « reproductible », tant au niveau individuel qu'au niveau d'un laboratoire[6]. La première raison touche au problème mentionné ci-dessus de la difficulté d'accès à l'ancienne littérature ; en matière d'analyse de données, une période de six mois peut déjà faire office de temps long et toute personne, à l'exception des plus méticuleuses dans la tenue de leur cahier de laboratoire, sait que reproduire une des *ses propres* figures après un tel délai peut parfois relever du casse-tête. La recherche reproductible ne va pas forcément faire disparaître instantanément les problèmes rencontrés dans ces circonstances, mais elle va permettre d'identifier leurs éventuelles sources – un changement de version d'un logiciel par exemple – de façon beaucoup plus rapide. Notre expérience d'une dizaine d'années avec ce type d'approches montre qu'elles apportent une bien plus grande pérennité au travail du chercheur. Ce qui vaut pour le chercheur « s'observant lui-même » à quelques mois ou années d'écart, vaut d'autant plus pour l'étudiant ou le stagiaire post-doctoral poursuivant le travail d'un de ces prédécesseurs, surtout si celui-ci a déjà quitté le laboratoire. Ainsi la recherche reproductible va automatiquement entraîner une conservation des savoir-faire et, par-là, faciliter leur transmission au sein d'une équipe, d'un laboratoire comme d'un institut.
Convaincu de l'intérêt de la recherche reproductible, le lecteur se demande sans doute comment la mettre en pratique. La recherche reproductible est depuis quelques années un domaine en plein développement et, comme tout domaine en pleine croissance, il se présente au novice, à travers la littérature, sous un jour assez chaotique. Le but de cet article, après avoir brièvement présenté le développement historique de la recherche reproductible, est de fournir une boussole, et une cartographie minimale, utiles au lecteur qui voudrait aller plus loin.

Une brève histoire de la recherche reproductible et de ses outils

La première tentative concrète de mise en œuvre d'« approches reproductibles », *au niveau des*

4. Ce qui implique de décrire cet environnement de façon suffisamment explicite.
5. Idéal qui nécessite – dans une certaine mesure au moins – la reproductibilité.
6. Si nous militons pour une « version forte » de la recherche reproductible comme mode de partage par défaut au sein d'une communauté scientifique, il nous semble important de souligner qu'un chercheur pourrait vouloir publier « comme avant » et néanmoins trouver une approche et des outils intéressants dans le présent article.

publications, est apparue en économie au début des années quatre-vingt (Dewald, Thursby, and Anderson 1986). Le *Journal of Money, Credit and Banking* a alors adopté une politique éditoriale demandant aux auteurs les programmes et données utilisés dans leurs articles « empiriques », ainsi que la mise à disposition de ceux-ci à tout chercheur sur simple demande. Cette mise à disposition s'est néanmoins faite de manière informelle par dépôt des codes et données dans un répertoire (d'ordinateur). Les approches reproductibles proposées par la suite peuvent être vues, en quelque sorte, comme le détournement (ou l'adaptation) d'outils créés dans un but assez différent. Ces outils sont ceux forgés par les informaticiens pour développer des logiciels fiables, bien documentés, faciles à faire évoluer et modifiables par d'autres personnes que leur auteur. Le premier outil est un moteur de production dont l'archétype dans le monde Unix est make : un logiciel, programmé par un langage de script, qui permet d'automatiser et d'ordonnancer la construction / compilation de logiciels « complexes » à partir de fichiers sources. Il est assez simple de remplacer le produit final, un logiciel complexe, par un article au format PDF (via LaTeX) et les compilations intermédiaires par des appels à des logiciels d'analyse de données en mode *batch* (non-interactif) – le résultat de tels appels étant par exemple la génération des figures de l'article. C'est l'idée utilisée par les géophysiciens du *Stanford Exploration Project*[7] (Claerbout and Karrenbach 1992). Au début des années 2000, des statisticiens, Friedrich Leisch et Tony Rossini (Leisch 2002b; Leisch 2002a; Leisch 2003; Rossini and Leisch 2003), se sont inspirés de la « programmation lettrée », proposée par Don Knuth lorsqu'il développait TeX (Knuth 1984). Ils ont ainsi créé la fonction Sweave du logiciel R qui traite un fichier au format texte (ASCII ou UTF-8) où le texte d'un article, écrit avec LaTeX, est mélangé aux lignes de code R qui génèrent les figures et les tables de l'article[8].

Un exemple

Nous avons préparé quelques versions – disponibles sur un dépôt GitHub associé à cet article – d'un cas concret de recherche reproductible dans un contexte qui devrait être proche d'un travail « quotidien » de statistique appliqué : téléchargement de données, chargement de celles-ci dans le logiciel d'analyse, vérifications de la fidélité de l'importation des données par génération de « résumés numériques », construction d'un graphe. Nous avons choisi la reproduction d'un graphe de William Playfair comme illustration. Nous mettons particulièrement l'accent sur la version combinant R et son extension R Markdown ainsi que sur la version combinant Python et le «carnet de notes» (notebook) IPython.

Conclusions

La recherche reproductible est bien plus qu'un *buzzword*, c'est une façon un peu différente de faire ce que le scientifique faisait déjà ; une façon de communiquer plus explicitement et de préserver son travail de façon plus « rationnelle » et plus systématique. C'est une approche encore incomplète dans la mesure où elle ne recouvre pas la phase de génération / collecte des données ; mais si ces dernières sont considérées comme « fixées » – ce qui est le contexte de travail typique des praticiens de la fouille de données –, elles doivent être partagées et deviennent ainsi critiquables et comparables. Il n'y a de cela que quelques années, dix ans tout au plus, pratiquer la recherche reproductible demandait un travail supplémentaire non négligeable au chercheur. Aujourd'hui, avec l'émergence de langages de programmation interactifs comme R et Python, avec des environnements de travail comme RStudio et IPython, avec la généralisation des langages de balisage légers comme pandoc markdown et des logiciels de gestion de version, changer ses habitudes pour rendre son travail quotidien reproductible ne demande guère plus qu'une demi-journée d'auto-formation. Les journaux scientifiques et

7. Le logiciel de recherche reproductible de ce groupe, Madagscar (Fomel and Hennenfent 2007), est maintenant basé sur le moteur de production scons.
8. Sweave recopie la partie texte du fichier d'entrée, telle quelle, dans un nouveau fichier LaTeX, exécute les lignes de codes puis inclut leurs résultats (figures et tables) dans le nouveau fichier.

les agences de financement sont, de leur coté, de plus en plus sensibles aux questions de partage des données et des codes ; les chercheurs ont donc tout intérêt à adopter les pratiques que nous venons d'exposer ; à tel point qu'il serait approprié, nous semble-t-il, de les inclure tôt dans les cursus universitaires. Alors ne ratez pas le train, il démarre maintenant et le prix de la place n'est vraiment pas élevé !

Références

Claerbout, Jon, and Martin Karrenbach. 1992. "Electronic Documents Give Reproducible Research a New Meaning." In *Proceedings of the 62nd Annual Meeting of the Society of Exploration Geophysics*, 601–4.

Collins, Francis S, and Tabak, Lawrence A. 2014 "NIH plans to enhance reproductibility" *Nature* 505 (7485): 612-613

Dewald, William G., Jerry G. Thursby, and Richard G. Anderson. 1986. "Replication in Empirical Economics: The Journal of Money, Credit, and Banking Project." *American Economic Review* 76 (4): 587–603.

Fomel, S., and G. Hennenfent. 2007. "Reproducible Computational Experiments Using Scons." In *Proc. IEEE Int'l Conf. Acoustics, Speech and Signal Processing*, 1257–60. 4.

Hinsen, Konrad. 2014. "Platforms for Publishing and Archiving Computer-Aided Research." F1000Research 3: 289. doi:10.12688/f1000research.5773.1.

Knuth, Donald E. 1984. "Literate Programming." *The Computer Journal* 27 (2): 97–111.

Leisch, Friedrich. 2002a. "Sweave: Dynamic Generation of Statistical Reports Using Literate Data Analysis." In *Compstat* 2002 — *Proceedings in Computational Statistics*, edited by Wolfgang Härdle and Bernd Rönz, 575–80. Physica Verlag, Heidelberg.

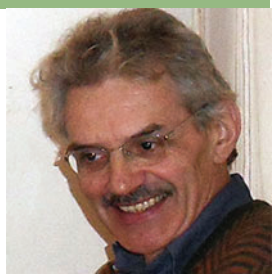
Leisch, Friedrich. 2002b. "Sweave, Part I: Mixing R and LaTeX." *R News* 2 (3): 28–31.

Leisch, Friedrich 2003. "Sweave, Part II: Package Vignettes." *R News* 3 (2): 21–24.

Polanyi, Michael. 1998. Personal Knowledge: Towards a Post-Critical Philosophy. Routledge.

Rossini, Anthony, and Friedrich Leisch. 2003. *Literate Statistical Practice*. UW Biostatistics Working Paper Series 194. University of Washington.

Études scientifiques : quelle validation ?
Compte rendu d'un Café de la statistique



Jean-François ROYER
SFdS

Comment s'assurer de la validité des résultats d'une étude scientifique ? Aujourd'hui la question est posée, après que de plus en plus d'articles publiés ont fait l'objet de rétractations. Des sites Internet[1] se consacrent à documenter les cas de retrait. Des statisticiens s'interrogent sur leurs procédures traditionnelles, et proposent de nouveaux standards. De façon plus large, le mouvement « pour une recherche reproductible » vise à permettre les validations multiples par des tiers. Ces réponses ne doivent pas dissimuler les causes des dérives, dont certaines sont à chercher dans les modes de financement et d'évaluation de la recherche.
Un Café de la Statistique s'est tenu sur ce thème en février 2015. L'intervenant était Stéphane Gregoir, Insee, ancien directeur de la recherche à l'EDHEC (Ecole des Hautes Etudes Commerciales) et ancien directeur du CREST (Centre de recherche en Economie et en Statistique). Vidéo et compte rendu détaillé sont disponibles sur le site de la SFdS

De plus en plus d'études scientifiques sont contestées après avoir été publiées. Le doute atteint beaucoup de disciplines. Les sciences biomédicales sont en première ligne : selon des chercheurs américains (Fang, Steen et Casadevall, 2012), deux mille articles publiés depuis 1975 dans des revues de bio-médecine et de sciences de la vie ont été retirés après leur publication, et le nombre de ces retraits est en forte augmentation depuis le début des années 2000. L'économie est également touchée : des articles d'économistes renommés, publiés dans des revues prestigieuses, ont dû être rectifiés après que des vérifications ont révélé que les conclusions étaient entachées d'erreurs[2] . Des exemples similaires se rencontrent dans d'autres disciplines, qu'elles relèvent des sciences « dures », y compris les mathématiques, ou des sciences humaines et sociales.

Un certain nombre de retraits sont dus à la découverte de plagiats, ou d'une version déjà publiée de la même étude. Les autres cas proviennent soit d'erreurs constatées après publication, soit de fraudes. La fraude ici consiste essentiellement en la construction de données fabriquées ou manipulées pour fournir un résultat désiré. Le plus souvent, c'est donc le comportement de l'auteur de l'article qui est en cause. Depuis l'assertion trop hâtive jusqu'à la fraude caractérisée, on constate toute une gamme de comportements susceptibles d'aboutir à la publication de conclusions fausses.

1. Comme le site « Retractation watch » cité dans « Des faussaires dans les labos », article de David Larousserie, Le Monde Sciences et technologie, 12 mai 2015
2. Un exemple souvent cité est (Reinhart & Rogoff 2011).

Un problème qui concerne la statistique

Les statisticiens sont particulièrement concernés par ce problème. Pourquoi ? Parce que leur discipline joue un rôle pivot dans la construction des preuves des résultats scientifiques. Bien rares sont les études empiriques dont les résultats peuvent être obtenus et présentés sans utiliser des tests statistiques. La validité de la preuve dépend du bon emploi de ces méthodes, standardisées il y a plusieurs décennies[3] . Trop souvent, certains auteurs appliquent mécaniquement des recettes en utilisant directement les sorties des logiciels sans en maîtriser les conditions d'utilisation : leur formation aux méthodes statistiques devrait être renforcée. Plus en profondeur, certains statisticiens réexaminent l'usage standard : les tailles d'échantillon requises sont-elles suffisantes ? Les niveaux d'erreur admis ne sont-ils pas trop grands ? En utilisant une approche bayésienne, des auteurs montrent comment en utilisant le seuil usuel (le fameux « alpha=5 % »), on est conduit à publier un pourcentage important d'études qui concluent à tort qu'un résultat est « significatif » (Johnson, 2013). Ces auteurs préconisent d'utiliser des standards beaucoup plus stricts pour établir les résultats scientifiques et restaurer la confiance du public dans la science.

Le fonctionnement de la recherche en question

D'autre part, des pratiques qui relèvent de l'organisation de la recherche « poussent à la faute ». Ainsi, les résultats négatifs sont peu considérés, rarement publiés, peu cités ; au contraire un résultat positif, même à la limite de la significativité, a des chances d'être souvent cité. Dans un contexte où la carrière d'un chercheur, et son accès aux sources de financement, dépendent pour une part du nombre de ses publications et du nombre des citations qui en sont faites, l'effet de ce biais peut être dramatique.
Le système des « referees », qui doivent examiner les articles avant publication dans les revues scientifiques, devrait constituer un garde-fou : un bon referee ne laisse pas passer un article contenant des conclusions fausses ou insuffisamment prouvées. Ce garde-fou ne fonctionne pas toujours correctement. Le « referee » dispose rarement d'une information complète : protocoles d'expérience, bases de données, programmes informatiques, tout cela ne lui est pas toujours communiqué, même si la situation évolue (voir plus loin). Et surtout, le referee ne travaille pas pour lui-même, mais pour la communauté scientifique : c'est un travail astreignant, peu visible et donc peu valorisé. Certains chercheurs confirmés acceptent ce rôle, mais, pris par les délais, sous-traitent les rapports demandés à des chercheurs plus jeunes, quand ce n'est pas à des étudiants insuffisamment expérimentés pour ce travail.
Enfin, lorsqu'un chercheur veut prouver « à tout prix » un résultat qui conforte une position générale qu'il a sur l'économie ou sur la société, il peut devenir moins « regardant » sur la rigueur de ses raisonnements, voire sur l'intégrité de ses données et de ses calculs.

La multiplication des mises en cause d'études scientifiques a provoqué une prise de conscience, et un appel à plus d'exigences dans l'établissement et la publication des résultats.
La sévérité à l'égard des fraudeurs se renforce. Aux Etats-Unis, s'agissant du domaine bio-médical, un « Office of research integrity » a été créé pour détecter la fraude et poursuivre les fraudeurs, et a déjà obtenu des condamnations.

La reproductibilité est-elle la solution ?

Mais la fraude n'est qu'une petite partie du problème, on l'a vu. Pour le traiter plus au fond, beaucoup considèrent qu'il faut privilégier l'effort vers une meilleure « reproductibilité » des recherches. L'échec rencontré par une tentative de reproduire un ensemble d'études

3. L'article de Neyman et Pearson sur les tests statistiques qui a fondé une pratique largement répandue date de 1933

fondamentales sur le cancer en 2012 a d'ailleurs beaucoup contribué à la prise de conscience récente (Begley, Ellis 2012). Si toutes les conditions sont remplies pour permettre à d'autres chercheurs de reproduire un travail, celui-ci se doit d'être irréprochable, sans quoi ses lacunes se verront très vite. Le mouvement « pour une recherche reproductible » prend de l'ampleur : de plus en plus de revues demandent aux chercheurs qui leur proposent des articles de fournir aussi leurs jeux de données et leurs programmes. Des logiciels spécifiques sont conçus pour aider les auteurs d'articles à satisfaire ces exigences[4].
Cela dit, tout le monde n'est pas d'accord sur le concept de reproductibilité, qui ne s'applique pas à toutes les formes de recherche scientifique[5]. Et sa mise en pratique se heurte à de nombreuses difficultés : par exemple, dans certains cas, les données confidentielles ne peuvent être communiquées à des tiers non identifiés, comme le sont les referees (anonymes).

En conclusion

L'enseignement d'une déontologie de la recherche, de plus en plus courante dans les écoles doctorales, permet d'informer les jeunes chercheurs sur l'éthique de leur profession.
Peut-être les incitations qui s'appliquent aux chercheurs en activité seraient-elles à revoir. Si ce sont ces incitations qui engendrent les dérives constatées, elles ne pourraient être contrecarrées que par des réformes de l'organisation même de la recherche, notamment en ce qui concerne le financement et l'évaluation des chercheurs.

Références

C. Glenn Begley, Lee M. Ellis, 2012 « Drug development: Raise standards for preclinical cancer research » Nature n°483
Ferric C. Fang, R. Grant Steen, Arturo Casadevall, 2012« Misconduct accounts for the majority of retracted scientific publications » Proceedings of the National Academy of Sciences - USA vol 109 n°42
John P.A. Ioannidis, 2005 « Why most published research findings are false » PLoS Medicine vol. 2 n°8
Valen E. Johnson, 2013 « Revised standards for statistical evidence » Proceedings of the National Academy of Sciences - USA n°1313476110
David Larousserie « Des faussaires dans les labos » Le Monde Sciences et technologie, 12 mai 2015
Jill Nelmark, 2015 « La guerre de la rétractation » Courrier International n°1269 du 26 février au 4 mars 2015
Carmen M. Reinhart and Kenneth S. Rogoff, « From Financial Crash to Debt Crisis », *The American Economic Review*, Vol. 101, No. 5 (Aug. 2011), pp. 1676-1706
« Unreliable research : Trouble in the lab » The Economist 19 octobre 2013
Compte rendu du Café de la statistique du 3 février 2015 « Comment s'assurer de la validité des arguments statistiques contenus dans les études scientifiques ?». Sur le site web de la SFdS, ce compte-rendu est accompagné de la vidéo de l'exposé de l'intervenant.
http://reproducibleresearch.net/ Site renfermant de nombreux onglets renvoyant eux- mêmes à des informations intéressantes sur la recherche reproductible, la lutte contre le plagiat et la fraude, etc...

4. Voir dans ce même numéro page 35 l'article de Christophe Pouzat, Andrew Davison et Konrad Hinsen
5. En histoire, sciences politiques, sociologie, géographie, la reproductibilité n'a pas de sens et d'autres critères de vérité ont été avancés.